the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **76**, 1057–1073 doi: 10.1111/tpj.12357

RESOURCE

Rice DB: an *Oryza* Information Portal linking annotation, subcellular location, function, expression, regulation, and evolutionary information for rice and Arabidopsis

Reena Narsai[1,2]*, James Devenish[1], Ian Castleden[2], Kabir Narsai[1], Lin Xu[1], Huixia Shou[3] and James Whelan[1]

[1]ARC Centre of Excellence in Plant Energy Biology, University of Western Australia, MCS Building M316, 35 Stirling Highway, Crawley 6009, Western Australia, Australia,

[2]Centre for Computational Systems Biology, University of Western Australia, MCS Building M316, 35 Stirling Highway, Crawley 6009, Western Australia, Australia, and

[3]State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

Received 26 June 2013; revised 30 September 2013; accepted 4 October 2013; published online 21 October 2013.

*For correspondence (e-mail reena.narsai@uwa.edu.au).

SUMMARY

Omics research in *Oryza sativa* (rice) relies on the use of multiple databases to obtain different types of information to define gene function. We present Rice DB, an *Oryza* information portal that is a functional genomics database, linking gene loci to comprehensive annotations, expression data and the subcellular location of encoded proteins. Rice DB has been designed to integrate the direct comparison of rice with Arabidopsis (*Arabidopsis thaliana*), based on orthology or 'expressology', thus using and combining available information from two pre-eminent plant models. To establish Rice DB, gene identifiers (more than 40 types) and annotations from a variety of sources were compiled, functional information based on large-scale and individual studies was manually collated, hundreds of microarrays were analysed to generate expression annotations, and the occurrences of potential functional regulatory motifs in promoter regions were calculated. A range of computational subcellular localization predictions were also run for all putative proteins encoded in the rice genome, and experimentally confirmed protein localizations have been collated, curated and linked to functional studies in rice. A single search box allows anything from gene identifiers (for rice and/or Arabidopsis), motif sequences, subcellular location, to keyword searches to be entered, with the capability of Boolean searches (such as AND/OR). To demonstrate the utility of Rice DB, several examples are presented including a rice mitochondrial proteome, which draws on a variety of sources for subcellular location data within Rice DB. Comparisons of subcellular location, functional annotations, as well as transcript expression in parallel with Arabidopsis reveals examples of conservation between rice and Arabidopsis, using Rice DB (http://ricedb.plantenergy.uwa.edu.au).

Keywords: *Oryza sativa*, rice, *Arabidopsis thaliana*, Arabidopsis, protein, subcellular location, transcript expression.

INTRODUCTION

The sequencing of the *Arabidopsis thaliana* (Arabidopsis) genome in 2000, followed by that of *Oryza sativa* (rice) and an increasing number of other species (Swarbreck *et al.*, 2008; Youens-Clark *et al.*, 2011) has stimulated efforts to define the functions of all genes in a plant genome, such as the Arabidopsis 2010 and RICE2020 projects, respectively (Chory *et al.*, 2000; Zhang *et al.*, 2008). Rice is a major food-producing crop and important monocot plant model (Han *et al.*, 2007). Thus, the completed rice genome sequence in 2005 enabled genome-wide approaches to be applied to rice research (Rice Genome, 2005). Functional annotation of genes depends to a large degree on various omic(s) approaches, such as transcriptomic, proteomic and/or metabolomic data sets used to provide insight under altered genetic/environmental conditions: for example in Arabidopsis (Borevitz and Ecker, 2004; Nordborg and Weigel, 2008). A good example of integrating traditional and post-genomic research in rice is the transcriptomic analysis of super-hybrid rice and its parents (Gibbs *et al.*, 2011). The advent of these omic data sets has led to the generation of various web-based public databases that give access to this data in ways that would be useful to scientists (Long *et al.*, 2008).

This flow of research from genomics to web-based public databases in plants is best seen for Arabidopsis, with more recent updates also including rice. A number of expression-based databases [the Bio-Array Resource (BAR; Schroder *et al.*, 2011) and Genevestigator (Wells *et al.*, 2010)], protein-based databases (ARAMEMNON; Schwacke *et al.*, 2003) and metabolite databases [Golm Metabolome Database (Kopka *et al.*, 2005) and Madison-Qingdao Metabolomics Consortium Database (Cui *et al.*, 2008)] are now available. A number of specialized databases are also available for Arabidopsis only, including the Arabidopsis Predicted Interactome (Geisler-Lee *et al.*, 2007) and the Subcellular Location Database for Arabidopsis (SUBA; Heazlewood *et al.*, 2005; Tanz *et al.*, 2013). A key element amongst these Arabidopsis databases is their integration via an international combined effort, the Multinational Arabidopsis Steering Committee, which provides an avenue to promote co-operative development and integration of resources. An example of this is the MASCP Gator database, which is a portal that draws on proteome data from a variety of databases for Arabidopsis (Joshi *et al.*, 2011). Also, the single major Arabidopsis database, The Arabidopsis Information Resource (TAIR), provides a place where data from various sources are combined (2005).

The importance of rice as a basic food source for approximately three billion people has led to a variety of independent post-genomic resources that have been applied in various studies (http://www.irri.org). However, unlike TAIR, there is not a single unified database for rice, with both the MSU Rice Genome Annotation Project (RGAP; Ouyang *et al.*, 2007) and the Rice Annotation Project Database (RAP-DB; Tanaka *et al.*, 2008) presenting mainly sequence and annotation information for all rice genes. Both of these databases are extremely useful for rice research, and have even incorporated new functions such as simple keyword searches and even gene expression information in RGAP. Also, both of these databases have now facilitated integration, by allowing the conversion of rice identifiers between these databases. Other useful databases for rice include specialized expression and co-expression databases such as the Rice Oligonucleotide Array Database (Jung *et al.*, 2008a), RiceXPro (Sato *et al.*, 2013b), Oryzaexpress (Hamada *et al.*, 2011), RiceFREND (Sato *et al.*, 2013a) and the Rice BAR-eFP browser (Toufighi *et al.*, 2005; Patel *et al.*, 2012), which all provide useful ways to examine transcript expression patterns in rice. Additionally, the Oryzabase database shows extensive rice information, curated by rice researchers (Kurata and Yamazaki, 2006). Another important rice resource is the Gramene database, which now also encompasses several genomes for rice and other species (Youens-Clark *et al.*, 2011). Gramene is very powerful for the comparison of genomes and genes using a variety of tools such as whole-genome alignments, evolutionary relationships of genes, synteny and genetic diversity, as well as linking to various resources for visualization of biochemical pathways (Jaiswal *et al.*, 2006). Similarly, the SALAD database (Mihara *et al.*, 2010) uses protein sequence information across several different plant species, including rice, to gain insight into protein function by revealing conserved protein motifs.

Despite the various specialized rice databases available, accessing and integrating these information sources is still challenging, especially because of the use of non-systematic identifiers. Furthermore, the outputs of these individual databases are often obtained in a piecemeal manner, requiring manual integration. Many desired data sets lack any direct, if any, references to subcellular location, which is arguably one of the most crucial pieces of information to determine the function of any protein (Koroleva *et al.*, 2005). The multicompartmental nature of eukaryotic cells allows specialization of function, which is reflected in the subcellular compartments of organelles. In Arabidopsis, the subcellular proteome is well advanced, with experimentally confirmed proteomes determined for the cell wall, plasma membrane, nucleus, plastid, mitochondria, Golgi, endoplasmic reticulum, peroxisome and cytosol (Agrawal *et al.*, 2010). In contrast, no comparably complete subcellular proteome map exists for rice to show how the subcellular proteomes of these two important plant models compare with each other. Additionally, there is currently no rice database that provides subcellular location information, in terms of collating the experimentally determined and predicted locations of rice proteins, as shown in SUBA, the subcellular location database for Arabidopsis (Heazlewood *et al.*, 2005; Tanz *et al.*, 2013). Finally, research to elucidate the function of rice genes, i.e. for predicted proteins, would benefit greatly from direct integration (and parallel comparison) of information on orthologous genes in Arabidopsis.

To address these issues, as well as provide completely new resources (including subcellular location information), a database for rice has been developed that integrates all of the above types of information. The rice database presented is a functional genomics database that performs integrated searches for functional annotations, subcellular location, expression levels, and putative or known regulatory elements, as well as showing orthologues between rice and Arabidopsis. Furthermore, the direct comparison with Arabidopsis in relation to a number of features including function and subcellular location facilitates greater understanding of the possible functions and locations of rice proteins, for which this information is unknown. Thus, Rice DB has been designed as a multilevel network that would be useful for researchers in various areas, from genomics and transcriptomics to proteomics. To present the ease of use and functions available in Rice DB, the rice mitochondrial proteome as well are other examples are shown.

RESULTS

Integrating information at a single website

For Arabidopsis, the TAIR database encompasses sequence information, function and expression annotations, bulk downloads and useful tools for gene/protein analysis (TAIR, 2005): however, whereas the existing rice databases do provide very useful and accurate information, it is not a simple task to move between these, and subcellular location information is almost completely lacking for rice. In Rice DB, a range of data types were collated and networked, combining substantial volumes of curated and computed data, generated in-house specifically for Rice DB (Figure 1a; Table 1).

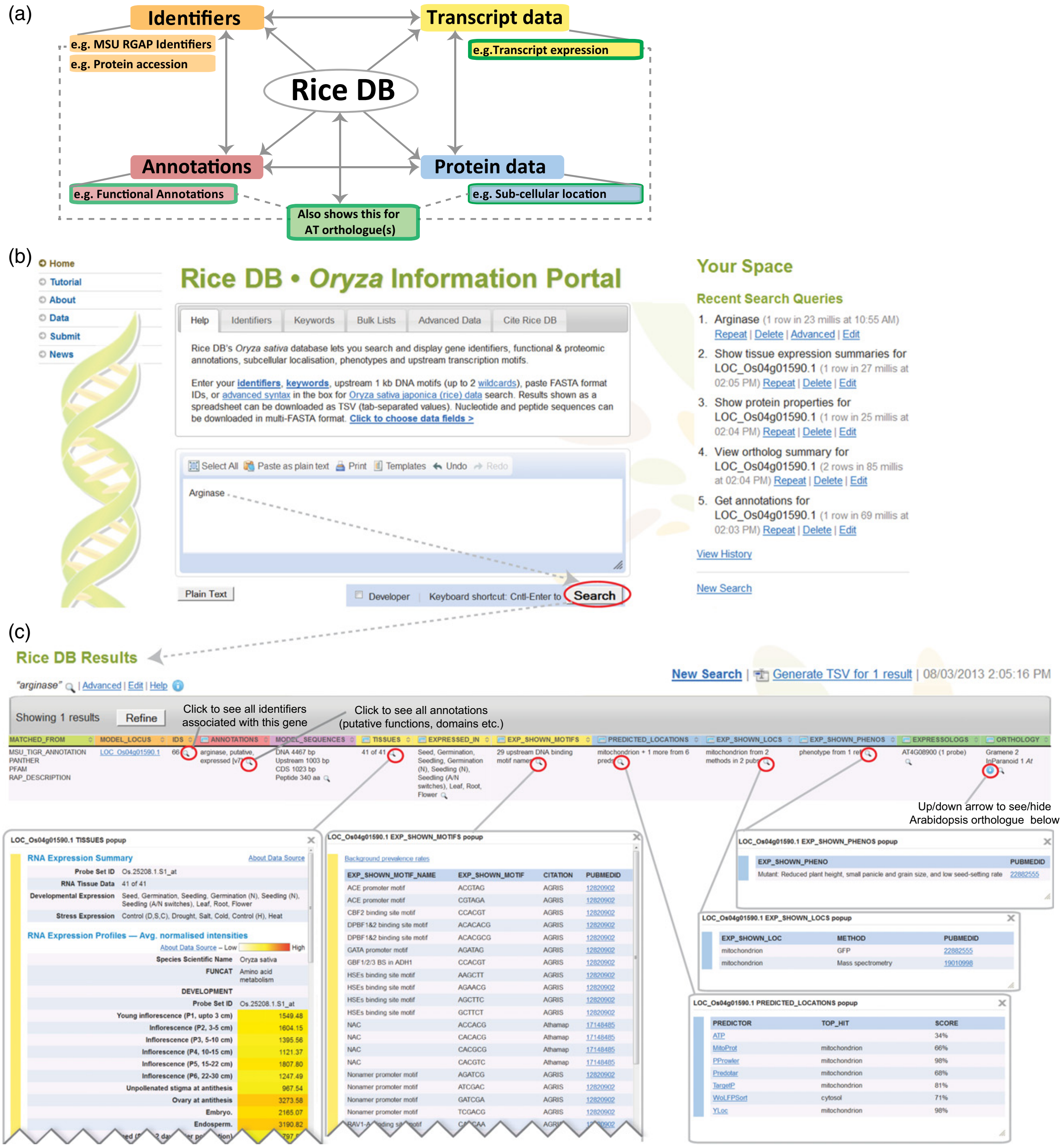
Firstly, it is immediately apparent that a range of identifier types are cited across different publications pertaining to rice, making it difficult to track a specific gene of interest. In Rice DB, different identifier types were collated from the numerous sources listed in Table 1, and any of these, including commonly used names (e.g. arginase) and gene symbols (e.g. OsARG) can be searched, and are converted to the commonly used Michigan State University (MSU) identifier. Given the multiple rice resources, there are also multiple putative function annotations for genes; thus putative function and domain annotations were collated and integrated in-house for Rice DB (Figure 1a; Table 1). Additionally, hundreds of microarrays were analysed in-house and expression information was collated to indicate whether or not a gene(s) is expressed (Figure 1a; Table 1). Furthermore, an extensive collection of protein information is presented in Rice DB, with protein properties, i.e. amino acid length, molecular weight and isoelectric point (calculated in-house; http://web.expasy.org/compute_pi), as well as predicted subcellular location (computed in-house; Table 1), manually compiled experimentally determined subcellular location and/or phenotype information (collated in-house from publications) listed.

Furthermore, unlike other rice databases, it is possible to simply view this information for Arabidopsis orthologues in parallel, where the Arabidopsis gene identifiers (AGIs), locus annotations (from TAIR), transcript expression (also analysed in–house), and predicted and experimentally determined location of Arabidopsis proteins (from SUBA; Heazlewood *et al.*, 2005; Tanz *et al.*, 2013) are also shown in Rice DB (indicated by the bold green borders of boxes in Figure 1a). Thus, users can begin with any data type in Rice DB, from genetic data (e.g. a specific promoter motif of interest) and transcript data (e.g. a differentially expressed gene list) to protein data (e.g. a protein list following mass spectrometry).

Oryza information portal: 'Google for rice'

As well as containing a wealth of information (Table 1), Rice DB presents a powerful 'Google'-style search engine, in that a wide range of searches are supported that can be entered into one search box. These include keywords (e.g. arginase; Figure 1b), a range of identifiers (e.g. LOC_ Os04g01590.1, Os04g106300, Q7X7N2, etc.), promoter motifs (e.g. AGATAG), domain annotations (e.g. Ureohydrolase), subcellular localization (e.g. mitochondria) and even AGIs (e.g. At4g08900.1). Note that the latter shows rice genes that are either orthologous or expressolog(s) to Arabidopsis genes, where an expressolog is defined on the basis of both orthology and conserved expression (Patel *et al.*, 2012). Boolean searches in Rice DB (such as AND/OR) also allow domain specialists to search and refine for specific fields, involving different combinations of information: such as annotation and protein location, e.g. 'kinase AND mitochondria in experimental location'; annotation and transcript expression, 'kinase AND expressed in seed'; subcellular location and expression. 'mitochondria in experimental location AND expressed in seed'. Thereby presenting a valuable feature that will be useful to researchers at any level. Note that although keyword searches within the MSU RGAP and RAP–DB are possible, and are very useful, these are limited to only finding keywords within the putative function annotations within each respective database. Furthermore, one of the intuitive features in Rice DB is that it detects typing errors and makes suggestions, e.g. if 'kina' or 'kinose' are searched, Rice DB will return with, 'Did you mean kinase?'

A recent study revealed the crucial role of a mitochondrial arginase (LOC_Os04g01590.1) in panicle development and grain production in rice (Ma *et al.*, 2012). This rice arginase, named OsARG, is used as an example gene to illustrate the functions and data available in Rice DB (Figure 1b, c; tutorial examples below the Rice DB search box). Upon searching for 'arginase' a summary for LOC_Os04g01590.1 is returned, using a systematic on-screen overview of the current gene research, with hyperlinks and side-by-side drill-down to scholarly data (Figure 1c). A colour-coded system was also created for Rice DB, to clarify the type of information being presented throughout the website (Figure 1c). There is a multitude of information at each of the individual levels (identifiers, annotations, etc.), thus only a short summary is shown after the search, with the option of viewing extended pop-up information by clicking on the magnifying glasses or opening the flat file by clicking the locus identifier (Figure 1c).


(a)
Identifiers
e.g. MSU RGAP Identifiers
e.g. Protein accession
Transcript data
e.g.Transcript expression
Rice DB
Annotations
e.g. Functional Annotations
Protein data
e.g. Sub-cellular location
Also shows this for AT orthologue(s)
(b)
(c)

Home
Tutorial
About
Data
Submit
News

Rice DB • *Oryza* Information Portal

Help | Identifiers | Keywords | Bulk Lists | Advanced Data | Cite Rice DB

Rice DB's *Oryza sativa* database lets you search and display gene identifiers, functional & proteomic annotations, subcellular localisation, phenotypes and upstream transcription motifs.

Enter your identifiers, keywords, upstream 1 kb DNA motifs (up to 2 wildcards), paste FASTA format IDs, or advanced syntax in the box for Oryza sativa japonica (rice) data search. Results shown as a spreadsheet can be downloaded as TSV (tab-separated values). Nucleotide and peptide sequences can be downloaded in multi-FASTA format. Click to choose data fields >

Select All | Paste as plain text | Print | Templates | Undo | Redo

Arginase

Plain Text

□ Developer | Keyboard shortcut: Cntl-Enter to Search

Your Space

Recent Search Queries

1. Arginase (1 row in 23 millis at 10:55 AM) Repeat | Delete | Advanced | Edit
2. Show tissue expression summaries for LOC_Os04g01590.1 (1 row in 27 millis at 02:05 PM) Repeat | Delete | Edit
3. Show protein properties for LOC_Os04g01590.1 (1 row in 25 millis at 02:04 PM) Repeat | Delete | Edit
4. View ortholog summary for LOC_Os04g01590.1 (2 rows in 85 millis at 02:04 PM) Repeat | Delete | Edit
5. Get annotations for LOC_Os04g01590.1 (1 row in 69 millis at 02:03 PM) Repeat | Delete | Edit

View History

New Search

Rice DB Results

"arginase" | Advanced | Edit | Help

New Search | Generate TSV for 1 result | 08/03/2013 2:05:16 PM

Showing 1 results | Refine

Click to see all identifiers associated with this gene

Click to see all annotations (putative functions, domains etc.)

MATCHED_FROM	MODEL_LOCUS	IDS	ANNOTATIONS	MODEL_SEQUENCES	TISSUES	EXPRESSED_IN	EXP_SHOWN_MOTIFS	PREDICTED_LOCATIONS	EXP_SHOWN_LOCS	EXP_SHOWN_PHENOS	EXPRESSOLOGS	ORTHOLOGY
MSU_TIGR_ANNOTATION PANTHER PFAM RAP_DESCRIPTION	LOC_Os04g01590.1	66	arginase, putative, expressed [v7]	DNA 4467 bp Upstream 1003 bp CDS 1023 bp Peptide 340 aa	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	29 upstream DNA binding motif names	mitochondrion + 1 more from 6 preds	mitochondrion from 2 methods in 2 pubs	phenotype from 1 ref	AT4G08900 (1 probe)	Gramene 2 InParanoid 1 *At*

Up/down arrow to see/hide Arabidopsis orthologue below

LOC_Os04g01590.1 TISSUES popup

RNA Expression Summary — About Data Source

Probe Set ID	Os.25208.1.S1_at
RNA Tissue Data	41 of 41
Developmental Expression	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower
Stress Expression	Control (D,S,C), Drought, Salt, Cold, Control (H), Heat

RNA Expression Profiles — Avg. normalised intensities

About Data Source – Low — High

Species Scientific Name	Oryza sativa
FUNCAT	Amino acid metabolism
DEVELOPMENT	
Probe Set ID	Os.25208.1.S1_at
Young inflorescence (P1, upto 3 cm)	1549.48
Inflorescence (P2, 3-5 cm)	1604.15
Inflorescence (P3, 5-10 cm)	1395.56
Inflorescence (P4, 10-15 cm)	1121.37
Inflorescence (P5, 15-22 cm)	1807.80
Inflorescence (P6, 22-30 cm)	1247.49
Unpollenated stigma at antithesis	967.54
Ovary at antithesis	3273.58
Embryo.	2165.07
Endosperm.	3190.82

LOC_Os04g01590.1 EXP_SHOWN_MOTIFS popup

Background prevalence rates

EXP_SHOWN_MOTIF_NAME	EXP_SHOWN_MOTIF	CITATION	PUBMEDID
ACE promoter motif	ACGTAG	AGRIS	12820902
ACE promoter motif	CGTAGA	AGRIS	12820902
CBF2 binding site motif	CCACGT	AGRIS	12820902
DPBF1&2 binding site motif	ACACACG	AGRIS	12820902
DPBF1&2 binding site motif	ACACGCG	AGRIS	12820902
GATA promoter motif	AGATAG	AGRIS	12820902
GBF1/2/3 BS in ADH1	CCACGT	AGRIS	12820902
HSEs binding site motif	AAGCTT	AGRIS	12820902
HSEs binding site motif	AGAACG	AGRIS	12820902
HSEs binding site motif	AGCTTC	AGRIS	12820902
HSEs binding site motif	GCTTCT	AGRIS	12820902
NAC	ACCACG	Athamap	17148485
NAC	CACACG	Athamap	17148485
NAC	CACGCG	Athamap	17148485
NAC	CACGTC	Athamap	17148485
Nonamer promoter motif	AGATCG	AGRIS	12820902
Nonamer promoter motif	ATCGAC	AGRIS	12820902
Nonamer promoter motif	GATCGA	AGRIS	12820902
Nonamer promoter motif	TCGACG	AGRIS	12820902

LOC_Os04g01590.1 EXP_SHOWN_PHENOS popup

EXP_SHOWN_PHENO	PUBMEDID
Mutant: Reduced plant height, small panicle and grain size, and low seed-setting rate	22882555

LOC_Os04g01590.1 EXP_SHOWN_LOCS popup

EXP_SHOWN_LOC	METHOD	PUBMEDID
mitochondrion	GFP	22882555
mitochondrion	Mass spectrometry	19010998

LOC_Os04g01590.1 PREDICTED_LOCATIONS popup

PREDICTOR	TOP_HIT	SCORE
ATP		34%
MitoProt	mitochondrion	66%
PProwler	mitochondrion	98%
Predotar	mitochondrion	68%
TargetP	mitochondrion	81%
WoLFPSort	cytosol	71%
YLoc	mitochondrion	98%

Figure 1. The user-friendly interface of Rice DB. (a) The major data types presented in Rice DB, showing the linked connections between them. The colour coding of each data type is maintained in the headings and side bars. (b) Screenshot showing the front page, where links to information about Rice DB, including the 'About', 'Data' and 'Tutorial' sections are shown on the left, next to a large search box allowing various entries. Note Boolean queries (AND/OR) are allowed. The right column shows examples of recent queries. The output after the search for arginase is shown here as a summary. Although 'arginase' returns one gene, note that when multiple genes are entered multiple rows are shown. (c) The summary output when 'arginase' was searched. Details of each data type are shown, with the coloured side bars representing the data type.

Although the sequencing and annotation of the rice genome has resulted in labelling all gene loci with systematic identifiers, many (or even the majority of) studies to date tend not to use any common or consistent identifier(s) in publications. This was particularly observed upon searching and compiling hundreds of publications for the purposes of including subcellular location and phenotype data to Rice DB. For example, one study of a luminal

Table 1 Outline of data presented in Rice DB

Data in Rice DB	Data subtype (brief description)	References
Alternative identifiers	MSU RGAP identifiers	Ouyang *et al.* (2007)
All identifiers were collated and matched to MSU identifiers; gene symbols were specifically curated and added into Rice DB	RAP-DB identifiers and descriptions	Tanaka *et al.* (2008)
	Gramene GenesDB symbol names[a]	Youens-Clark *et al.* (2011), and curated for Rice DB
	Probe set IDs – Rice Oligo. Array DB (ROAD)	Jung *et al.* (2008b)
	Oryzabase (inc. RefSeq, Unigene etc.)	Kurata and Yamazaki (2006)
	Arabidopsis - For Arabidopsis, AGIs were used as identifiers (Swarbreck *et al.*, 2008)	
Annotations	MSU putative function annotation	Ouyang *et al.* (2007)
Annotations were compiled to enable keyword searches across any of the sources.	RAP gene description	Tanaka *et al.* (2008)
	Genebin (Funcats at ANU)[a]	Goffard and Weiller (2007), and curated for Rice DB
	GO slim: ontology, domain (C/F/P)	Ouyang *et al.* (2007)
	Domain annotations	
	Coil predictions	http://www.ebi.ac.uk/Tools/pfa/iprscan/
	FingerPRINTScan	Attwood *et al.* (2012)
	Gene3D	http://gene3d.biochem.ucl.ac.uk/
	InterPro	http://www.ebi.ac.uk/interpro/
	PROSITE	Sigrist *et al.* (2010)
	Panther	Thomas *et al.* (2003)
	Pfam	Punta *et al.* (2012)
	SMART	Letunic *et al.* (2012)
	SuperFamily	Madera *et al.* (2004)
	TMHMM	http://www.cbs.dtu.dk/services/TMHMM/
	Arabidopsis: annotations were derived from TAIR (Swarbreck *et al.*, 2008)	
Transcript data	Transcript expression: data from multiple sources (see Experimental procedures)	
Microarray data were normalized and expression annotations were generated for Rice DB. The occurrence of all possible hexamers and *cis*-acting elements known to function in regulating expression has been calculated and shown. Known miRNA targets have also been matched and annotated.	RNA tissue data (no. tissues expressed in)	Analysed and compiled for Rice DB
	Expressed in ('expression annotation')	Analysed, compiled and annotated for Rice DB
	Stress expression ('expression annotation')	Analysed, compiled and annotated for Rice DB
	Experimentally confirmed DNA binding motifs: matched in 1 kb upstream regions, occurrences calculated	
	Motifs/CAREs known to be functional	AGRIS, Athamap, matched in rice promoters
	miRNAs: known miRNA targets and sequences presented in miRNA *et al.* are annotated	
	miRNA target and sequence information	Jeong *et al.* (2011)
	Arabidopsis: transcript expression data also shown (see Experimental procedures)	
Protein data	Predicted locations: output from each predictor is presented for Rice DB	
Peptide sequences for all the encoded genes in the rice genome were run through each computational predictor, and outputs are presented in Rice DB. In addition, published literature was searched and compiled showing experimentally determined localization. For those with experimentally determined localization, the phenotype is indicated if one was determined upon genetic alteration.	Ambiguous targeting predictor	Mitschke *et al.* (2009)
	ChloroP location predictor	Emanuelsson *et al.* (2007)
	MitoProt 2 location predictor	Claros (1995)
	PredictNLS location predictor	https://rostlab.org/owiki/index.php/PredictNLS
	Predotar location predictor	Small *et al.* (2004)
	ProteinProwler location predictor	Hawkins and Boden (2006)
	PTS1 location predictor	Neuberger *et al.* (2003)
	SignalP location predictor	Emanuelsson *et al.* (2007)
	TargetP location predictor	Emanuelsson *et al.* (2007)
	WoLFPSort location predictor	Horton *et al.* (2007)
	YLoc location predictor	Briesemeister *et al.* (2010)
	Experimentally confirmed locations: compiled for Rice DB	
	Experimentally confirmed phenotypes: compiled for Rice DB	
	Arabidopsis: predicted and experimental locations shown from SUBA (Heazlewood *et al.*, 2005)	

(continued)

Table 1 (continued)

Data in Rice DB	Data subtype (brief description)	References
Orthology with *Arabidopsis thaliana*		
Orthology data were matched and compiled for Rice DB from Inparanoid, Gramene and Expressologs.	Inparanoid orthology: clusters of orthologous genes with scores are shown	(Ostlund *et al.*, 2010)
	Gramene orthology: sequence identity between genes is shown	(Youens-Clark *et al.*, 2011)
	Expressologs: closest Expressolog gene is indicated	(Patel *et al.*, 2012)

[a] Manual amendment or addition to this category for Rice DB. Details of all these sources, including web links, are also shown on the Data page in Rice DB.

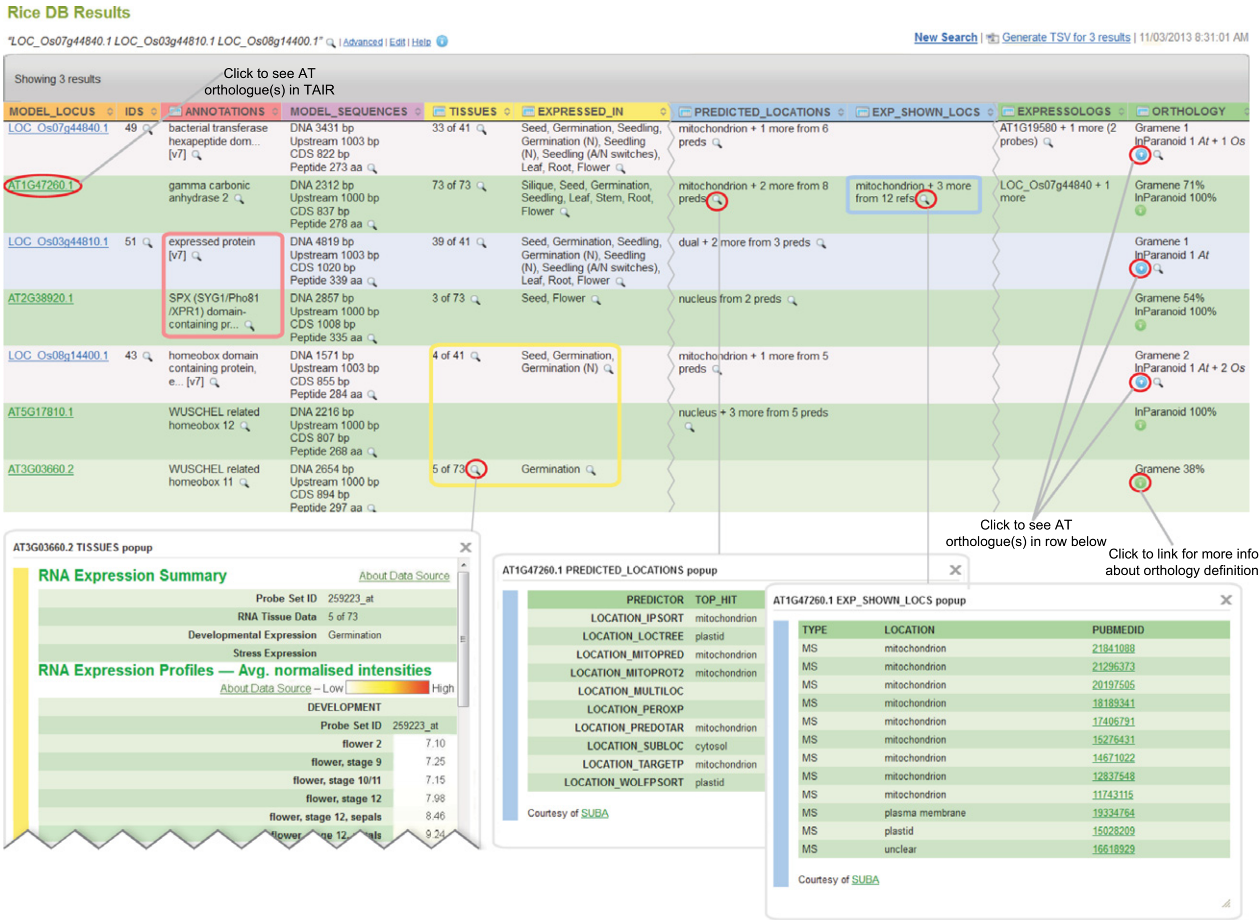
binding protein named it BiP, and AK065743 was given as the refseq accession (Yasuda *et al.*, 2009), whereas the same protein was name BiP3 and the MSU identifier LOC_Os02g02410.1 was given in another study (Park *et al.*, 2010). To discover whether these referred to the same gene/protein, AK065743 was first searched in the National Center for Biotechnology Information (NCBI) database, which showed that this refers to Os02g0115900 (a Rice Annotation Project (RAP) identifier), and this was then converted to an MSU identifier using either the MSU or RAP database to reveal that it also converts to LOC_Os02g02410.1. This was only one of several examples of where this happened, and this lack of consistency and difficulty in identifying genes/proteins can even, and has, led to different groups characterising the same protein. Thus, in Rice DB, alternative gene/protein/microarray identifiers are all automatically converted to the standard MSU identifier (LOC_OsXgXXXXX.X), and all identifiers can be viewed in parallel (orange column; Figure 1c). This conversion is a prerequisite before any integration of data sets can be routinely achieved.

Thus, a search using the gene symbol 'OsARG' or protein accessions 'Os04g106300' or 'Q7X7N2' also displays the same summary after conversion to the same MSU identifier 'LOC_Os04g01590.1' (as seen in Figure 1c). Details of each data type shown in the summary can also quickly be viewed by clicking on the magnifying glasses (Figure 1c), or in the case of multiple genes, by clicking on the spreadsheet icons to the left of the column headings. For example, the 'tissues' column shows that the arginase (LOC_Os04g01590.1) is expressed in 41 out of 41 analysed tissues, and clicking on the magnifying glass shows a pop-up box with the expression intensities indicated as a heat map (yellow sidebar; Figure 1c). Similarly, for the experimentally confirmed motifs, the name, sequence and hyperlinked PubMed identifier linking to the resource presenting this motif is shown (pop-up with yellow sidebar; Figure 1c). Three columns are shown for protein information, and the pop-up windows with the blue side bars show the details of these (Figure 1c). Firstly, the output from the different protein sublocation predictors are shown, e.g. when the arginase protein sequence was analysed by a range of subcellular location predictors, five out of the six predicted a mitochondrial location (e.g. Mitoprot, PProwler, etc; Figure 1c). Secondly, the experimentally demonstrated protein localization is shown, as well as the method used and a link(s) to the publication confirming this location. For this arginase, a mitochondrial location was shown in two publications: one of these was by green fluorescent protein tagging (GFP); the other study carried out organelle isolation followed by mass spectrometry to assign a mitochondrial location (Figure 1c). Lastly, a short description of the phenotype is shown for proteins with experimentally confirmed subcellular location(s). For example, it is shown that mutation in this arginase gene results in reduced plant height, and small panicle and grain size (Figure 1c).

Although these pop-ups are useful for quick checks, clicking on the hyperlinked model locus (orange column; Figure 1c) opens a flat file, which first repeats the 'Summary' at the top of the page and then follows with details of identifiers, annotations, transcript data, protein data, orthology and expressology associated with that gene (details in Table 1, and also on the 'Data' page at the Rice DB website). Outputs from Rice DB have been designed for research purposes; therefore, a separate line for each gene is shown when multiple hits occur, and all information can be readily downloaded as a tab-delimited file ('Generate TSV' icon above grey bar; Figure 1c). Details on how to use Rice DB are also shown in the interactive 'Tutorials' on the Rice DB webpage (left panel; Figure 1b).

Using Arabidopsis orthology to gain insight into rice

Despite the variety of compiled and newly generated data for rice in Rice DB, there are still hundreds, if not thousands, of genes/proteins for which specific information cannot be found. To address this problem, Rice DB allows simple, parallel access to Arabidopsis data. To date,



Rice DB Results

"LOC_Os07g44840.1 LOC_Os03g44810.1 LOC_Os08g14400.1" | Advanced | Edit | Help

New Search | Generate TSV for 3 results | 11/03/2013 8:31:01 AM

Showing 3 results

Click to see AT orthologue(s) in TAIR

MODEL_LOCUS	IDS	ANNOTATIONS	MODEL_SEQUENCES	TISSUES	EXPRESSED_IN	PREDICTED_LOCATIONS	EXP_SHOWN_LOCS	EXPRESSOLOGS	ORTHOLOGY
LOC_Os07g44840.1	49	bacterial transferase hexapeptide dom... [v7]	DNA 3431 bp Upstream 1003 bp CDS 822 bp Peptide 273 aa	33 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	mitochondrion + 1 more from 6 preds		AT1G19580 + 1 more (2 probes)	Gramene 1 InParanoid 1 *At* + 1 *Os*
AT1G47260.1		gamma carbonic anhydrase 2	DNA 2312 bp Upstream 1000 bp CDS 837 bp Peptide 278 aa	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	mitochondrion + 2 more from 8 preds	mitochondrion + 3 more from 12 refs	LOC_Os07g44840 + 1 more	Gramene 71% InParanoid 100%
LOC_Os03g44810.1	51	expressed protein [v7]	DNA 4819 bp Upstream 1003 bp CDS 1020 bp Peptide 339 aa	39 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	dual + 2 more from 3 preds			Gramene 1 InParanoid 1 *At*
AT2G38920.1		SPX (SYG1/Pho81 /XPR1) domain-containing pr...	DNA 2857 bp Upstream 1000 bp CDS 1008 bp Peptide 335 aa	3 of 73	Seed, Flower	nucleus from 2 preds			Gramene 54% InParanoid 100%
LOC_Os08g14400.1	43	homeobox domain containing protein, e... [v7]	DNA 1571 bp Upstream 1003 bp CDS 855 bp Peptide 284 aa	4 of 41	Seed, Germination, Germination (N)	mitochondrion + 1 more from 5 preds			Gramene 2 InParanoid 1 *At* + 2 *Os*
AT5G17810.1		WUSCHEL related homeobox 12	DNA 2216 bp Upstream 1000 bp CDS 807 bp Peptide 268 aa			nucleus + 3 more from 5 preds			InParanoid 100%
AT3G03660.2		WUSCHEL related homeobox 11	DNA 2654 bp Upstream 1000 bp CDS 894 bp Peptide 297 aa	5 of 73	Germination				Gramene 38%

Click to see AT orthologue(s) in row below

Click to link for more info. about orthology definitions

AT3G03660.2 TISSUES popup

RNA Expression Summary

About Data Source

Probe Set ID	259223_at
RNA Tissue Data	5 of 73
Developmental Expression	Germination
Stress Expression	

RNA Expression Profiles — Avg. normalised intensities

About Data Source – Low High

DEVELOPMENT	
Probe Set ID	259223_at
flower 2	7.10
flower, stage 9	7.25
flower, stage 10/11	7.15
flower, stage 12	7.98
flower, stage 12, sepals	8.46
flower [illegible] ge 12, [illegible] als	9.24

AT1G47260.1 PREDICTED_LOCATIONS popup

PREDICTOR	TOP_HIT
LOCATION_IPSORT	mitochondrion
LOCATION_LOCTREE	plastid
LOCATION_MITOPRED	mitochondrion
LOCATION_MITOPROT2	mitochondrion
LOCATION_MULTILOC	
LOCATION_PEROXP	
LOCATION_PREDOTAR	mitochondrion
LOCATION_SUBLOC	cytosol
LOCATION_TARGETP	mitochondrion
LOCATION_WOLFPSORT	plastid

Courtesy of SUBA

AT1G47260.1 EXP_SHOWN_LOCS popup

TYPE	LOCATION	PUBMEDID
MS	mitochondrion	21841088
MS	mitochondrion	21296373
MS	mitochondrion	20197505
MS	mitochondrion	18189341
MS	mitochondrion	17406791
MS	mitochondrion	15276431
MS	mitochondrion	14671022
MS	mitochondrion	12837548
MS	mitochondrion	11743115
MS	plasma membrane	19334764
MS	plastid	15028209
MS	unclear	16618929

Courtesy of SUBA

Figure 2. An example of the output after three genes/proteins in rice were searched in Rice DB. After clicking on the down arrow present in the orthology column, it is possible to see parallel information for the orthologous gene(s) in Arabidopsis within the Rice DB output table. Examples demonstrating the usefulness of showing Arabidopsis gene descriptions, expression annotations and subcellular locations in parallel are shown (in the pink, yellow and blue boxes, respectively). Examples of the pop-up windows are also shown for the expression and protein subcellular location(s) data.

numerous databases allow specific analyses of both Arabidopsis and rice, such as Gramene (Youens-Clark *et al.*, 2011) and the MIPS PlantDB (Nussbaumer *et al.*, 2013), which allow comparisons of genetic synteny, whereas databases such as ATTEDII (Obayashi *et al.*, 2011) and OryzaExpress (Hamada *et al.*, 2011) facilitate co-expression analysis. Similarly, the SALAD database (Mihara *et al.*, 2010) and PRIN database (Gu *et al.*, 2011) facilitate protein motif and interaction analysis in both rice and Arabidopsis. However, whereas some of these databases do allow a comparison of results between both species, others only facilitate separate analysis of both species, but under the same conditions, i.e. within the same database.

In Rice DB, multiple levels of data can easily be compared between Arabidopsis and rice, from transcript expression to protein properties and subcellular localization, all within the same database for both species (Figure 2). This is achieved through orthology, based on two different methods including sequence identity (as computed in Gramene; Youens-Clark *et al.*, 2011) as well as InParanoid (Ostlund *et al.*, 2010). Furthermore, a recent study has used co-expression and orthology to generate 'Expressologs' between rice and Arabidopsis genes (Patel *et al.*, 2012), and these are also shown in Rice DB (Figure 2). Thus, the user can chose to employ one or more of these preferred methodologies.

In the Rice DB summary output, the last two columns show the Arabidopsis expressologs and orthologues (Figures 1c and 2). Clicking the blue arrow in this column drops one (or more) rows below to reveal parallel information for the Arabidopsis orthologues(s) (rows shaded in green; Figure 2). More information about the orthology can be gained by clicking the green information icon (Figure 2). To demonstrate the ease and biological usefulness of showing Arabidopsis information in parallel, three example rice genes were searched (Figure 2). The first protein (encoded by LOC_Os07g44840.1) is annotated as a bacterial transferase hexapeptide domain-containing protein, which is predicted to be mitochondrial, but there is no experimental confirmation of the subcellular location (Figure 2). By using the Arabidopsis orthology link, it is easily shown that this protein has an Arabidopsis orthologue (At1g47260.1) of similar length, with 71% sequence identity, based on Gramene (Youens-Clark *et al.*, 2011), and with 100% confidence of orthology, based on InParanoid (Ostlund *et al.*, 2010) (green row below LOC_Os07g44840.1; Figure 2). However, in contrast to its rice orthologue, the subcellular location of At1g47260.1 has been experimentally confirmed to be mitochondrial, based on evidence from 12 different publications (blue box; Figure 2). The experimental method(s) used and links to publications are also shown (PubMed identifiers; green pop-up window; Figure 2). Specifically, this Arabidopsis protein (At1g47260.1) has been shown to be mitochondrial and part of complex I in Arabidopsis (Meyer *et al.*, 2008; Klodmann *et al.*, 2011), and clicking on At1g47260.1 (circled red; Figure 2) opens up the TAIR page for this gene directly, where other publications relating to this protein are also shown below. Given the high sequence identity and conserved, essential role of complex I in the electron transport chain, knowledge of the mitochondrial location of the protein encoded by At1g47260.1 sheds light onto the likely subcellular location of LOC_Os07g44840.1. In this way, Rice DB provides novel insight into the subcellular location of rice proteins, by presenting and linking to the known subcellular location information of Arabidopsis orthologues (for which this is collated in SUBA; Heazlewood *et al.*, 2003; Tanz *et al.*, 2013). Thus, Rice DB allows researchers to easily gain insight into the subcellular location for hundreds of rice proteins with Arabidopsis orthologues, for which this information is already known.

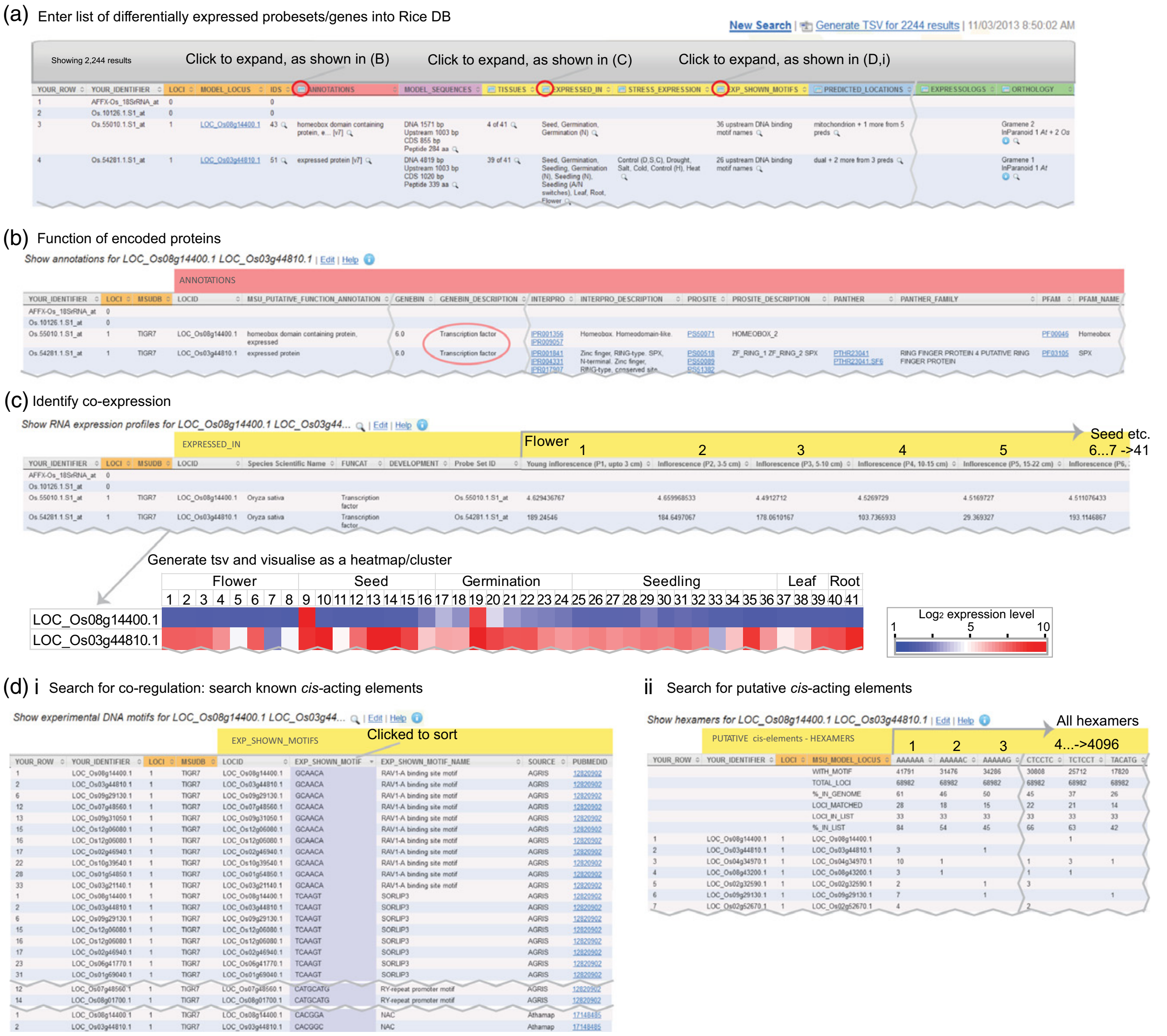
Similarly, although many rice genes lack detailed functional annotation, often the Arabidopsis orthologues are more informatively annotated. For example, there are nearly 14 300 rice genes annotated as ‘expressed protein’ in the MSU database, with many having no further (functional) description. An example is shown for LOC_Os03g44810.1, where the MSU annotation calls this an ‘expressed protein’, whereas the Arabidopsis orthologue (At2g38920.1) is annotated as an SPX domain-containing protein in TAIR (pink box; Figure 2). This annotation was also reinforced when the compiled domain annotations in Rice DB also showed that LOC_Os03g44810.1 does in fact contain an SPX domain, according to Interpro, Prosite and Pfam. Thus, having parallel annotation information for Arabidopsis orthologues can be useful for rice proteins where little functional annotation information is shown.

Lastly, comparing transcript expression between Arabidopsis and rice can also provide functional insight. For example, LOC_Os08g14400.1 is most highly expressed in seed/germination rice, with expression shown in four out of 41 tissues (yellow box; Figure 2). Using the orthology information in Rice DB, we can easily show that this expression is conserved in Arabidopsis, where At3g03660.2 is also expressed specifically during germination, in five out of the 73 tissues (yellow box; green pop-up window; Figure 2). This example was taken from a rice germination study (Howell *et al.*, 2009), in which this gene was shown to be highly expressed during germination, and this was also conserved for the Arabidopsis orthologue using the BAR efp browser (Toufighi *et al.*, 2005). Although, this was only done for a few genes during germination in that study (Howell *et al.*, 2009), it demonstrates how any gene or set of genes can be easily searched in Rice DB to examine and reveal conservation in expression patterns between Arabidopsis and rice. The usefulness of combining these data is also demonstrated in the recently updated rice BAR efp browser, showing expressologs combining expression and orthology as a highly informative way of revealing the conservation or divergence between species (Patel *et al.*, 2012). As indicated in that study, it is also important to point out here that there are also significant differences between Arabidopsis and rice, or dicots and monocots in general, that must be considered when interpreting these comparisons (Narsai *et al.*, 2010; Patel *et al.*, 2012). Details of these are described on the ‘Combined Orthology Summary’ page of the ‘Data’ pages.

Expression analysis for rice genes

In TAIR, when an Arabidopsis gene/protein is searched, the flat file shows annotation and function information as well as expression annotations: e.g. expressed in ‘seed, germination’. These expression annotations show the tissues/developmental stages in which the specific gene of interest is expressed. This is not only useful for transcriptomic studies, but even for proteomic research, where it can be useful to distinguish protein isoforms based on transcript expression patterns. To our knowledge, this is currently not found in other rice databases, and thus hundreds of microarrays for rice (and Arabidopsis) were analysed in parallel to generate expression annotations in Rice DB (Figure 3a).

During the course of omics research, often a list of genes/proteins are identified. A transcriptomic analysis workflow using Rice DB is presented in Figure 3. For example, following a microarray study a set of differentially expressed genes can be entered into Rice DB (Figure 3a). Given that Rice DB accepts Affymetrix microarray probe sets, identifier conversion was not necessary (identifiers retrieved from ROAD; Jung *et al.*, 2008b; Table 1). As an example, the 2244 differentially expressed probe sets identified during germination (Howell *et al.*, 2009) were entered into Rice DB and the summary for the first four probe sets are shown (Figure 3a). The first two probe sets do not match to specific MSU gene identifiers and therefore these


(a) Enter list of differentially expressed probesets/genes into Rice DB
Click to expand, as shown in (B)
Click to expand, as shown in (C)
Click to expand, as shown in (D,i)

New Search | Generate TSV for 2244 results | 11/03/2013 8:50:02 AM

Showing 2,244 results

YOUR_ROW	YOUR_IDENTIFIER	LOCI	MODEL_LOCUS	IDS	ANNOTATIONS	MODEL_SEQUENCES	TISSUES	EXPRESSED_IN	STRESS_EXPRESSION	EXP_SHOWN_MOTIFS	PREDICTED_LOCATIONS	EXPRESSOLOGS	ORTHOLOGY
1	AFFX-Os_18SrRNA_at	0		0									
2	Os.10126.1.S1_at	0		0									
3	Os.55010.1.S1_at	1	LOC_Os08g14400.1	43	homeobox domain containing protein, e... [v7]	DNA 1571 bp Upstream 1003 bp CDS 855 bp Peptide 284 aa	4 of 41	Seed, Germination, Germination (N)		36 upstream DNA binding motif names	mitochondrion + 1 more from 5 preds		Gramene 2 InParanoid 1 *At* + 2 *Os*
4	Os.54281.1.S1_at	1	LOC_Os03g44810.1	51	expressed protein [v7]	DNA 4819 bp Upstream 1003 bp CDS 1020 bp Peptide 339 aa	39 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	Control (D,S,C), Drought, Salt, Cold, Control (H), Heat	26 upstream DNA binding motif names	dual + 2 more from 3 preds		Gramene 1 InParanoid 1 *At*

Show annotations for LOC_Os08g14400.1 LOC_Os03g44810.1 | Edit | Help

ANNOTATIONS

YOUR_IDENTIFIER	LOCI	MSUDB	LOCID	MSU_PUTATIVE_FUNCTION_ANNOTATION	GENEBIN	GENEBIN_DESCRIPTION	INTERPRO	INTERPRO_DESCRIPTION	PROSITE	PROSITE_DESCRIPTION	PANTHER	PANTHER_FAMILY	PFAM	PFAM_NAME
AFFX-Os_18SrRNA_at	0													
Os.10126.1.S1_at	0													
Os.55010.1.S1_at	1	TIGR7	LOC_Os08g14400.1	homeobox domain containing protein, expressed	6.0	Transcription factor	IPR001356 IPR009057	Homeobox. Homeodomain-like.	PS50071	HOMEOBOX_2			PF00046	Homeobox
Os.54281.1.S1_at	1	TIGR7	LOC_Os03g44810.1	expressed protein	6.0	Transcription factor	IPR001841 IPR004331 IPR017907	Zinc finger, RING-type. SPX, N-terminal. Zinc finger, RING-type, conserved site.	PS00518 PS50089 PS51382	ZF_RING_1 ZF_RING_2 SPX	PTHR23041 PTHR23041:SF6	RING FINGER PROTEIN 4 PUTATIVE RING FINGER PROTEIN	PF03105	SPX

Show RNA expression profiles for LOC_Os08g14400.1 LOC_Os03g44... | Edit | Help

EXPRESSED_IN

YOUR_IDENTIFIER	LOCI	MSUDB	LOCID	Species Scientific Name	FUNCAT	DEVELOPMENT	Probe Set ID	Young inflorescence (P1, upto 3 cm)	Inflorescence (P2, 3-5 cm)	Inflorescence (P3, 5-10 cm)	Inflorescence (P4, 10-15 cm)	Inflorescence (P5, 15-22 cm)	Inflorescence (P6,
AFFX-Os_18SrRNA_at	0												
Os.10126.1.S1_at	0												
Os.55010.1.S1_at	1	TIGR7	LOC_Os08g14400.1	Oryza sativa	Transcription factor		Os.55010.1.S1_at	4.629436767	4.659968533	4.4912712	4.5269729	4.5169727	4.511076433
Os.54281.1.S1_at	1	TIGR7	LOC_Os03g44810.1	Oryza sativa	Transcription factor		Os.54281.1.S1_at	189.24546	184.6497067	178.0610167	103.7365933	29.369327	193.1146867

(d) i Search for co-regulation: search known *cis*-acting elements
Clicked to sort

Show experimental DNA motifs for LOC_Os08g14400.1 LOC_Os03g44... | Edit | Help

EXP_SHOWN_MOTIFS

YOUR_ROW	YOUR_IDENTIFIER	LOCI	MSUDB	LOCID	EXP_SHOWN_MOTIF	EXP_SHOWN_MOTIF_NAME	SOURCE	PUBMEDID
1	LOC_Os08g14400.1	1	TIGR7	LOC_Os08g14400.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
2	LOC_Os03g44810.1	1	TIGR7	LOC_Os03g44810.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
6	LOC_Os09g29130.1	1	TIGR7	LOC_Os09g29130.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
12	LOC_Os07g48560.1	1	TIGR7	LOC_Os07g48560.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
13	LOC_Os09g31050.1	1	TIGR7	LOC_Os09g31050.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
15	LOC_Os12g06080.1	1	TIGR7	LOC_Os12g06080.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
16	LOC_Os12g06080.1	1	TIGR7	LOC_Os12g06080.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
17	LOC_Os02g46940.1	1	TIGR7	LOC_Os02g46940.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
22	LOC_Os10g39540.1	1	TIGR7	LOC_Os10g39540.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
28	LOC_Os01g54850.1	1	TIGR7	LOC_Os01g54850.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
33	LOC_Os03g21140.1	1	TIGR7	LOC_Os03g21140.1	GCAACA	RAV1-A binding site motif	AGRIS	12820902
1	LOC_Os08g14400.1	1	TIGR7	LOC_Os08g14400.1	TCAAGT	SORLIP3	AGRIS	12820902
2	LOC_Os03g44810.1	1	TIGR7	LOC_Os03g44810.1	TCAAGT	SORLIP3	AGRIS	12820902
6	LOC_Os09g29130.1	1	TIGR7	LOC_Os09g29130.1	TCAAGT	SORLIP3	AGRIS	12820902
15	LOC_Os12g06080.1	1	TIGR7	LOC_Os12g06080.1	TCAAGT	SORLIP3	AGRIS	12820902
16	LOC_Os12g06080.1	1	TIGR7	LOC_Os12g06080.1	TCAAGT	SORLIP3	AGRIS	12820902
17	LOC_Os02g46940.1	1	TIGR7	LOC_Os02g46940.1	TCAAGT	SORLIP3	AGRIS	12820902
23	LOC_Os06g41770.1	1	TIGR7	LOC_Os06g41770.1	TCAAGT	SORLIP3	AGRIS	12820902
31	LOC_Os01g69040.1	1	TIGR7	LOC_Os01g69040.1	TCAAGT	SORLIP3	AGRIS	12820902
12	LOC_Os07g48560.1	1	TIGR7	LOC_Os07g48560.1	CATGCATG	RY-repeat promoter motif	AGRIS	12820902
14	LOC_Os08g01700.1	1	TIGR7	LOC_Os08g01700.1	CATGCATG	RY-repeat promoter motif	AGRIS	12820902
1	LOC_Os08g14400.1	1	TIGR7	LOC_Os08g14400.1	CACGGA	NAC	Athamap	17148485
2	LOC_Os03g44810.1	1	TIGR7	LOC_Os03g44810.1	CACGGC	NAC	Athamap	17148485

Show hexamers for LOC_Os08g14400.1 LOC_Os03g44810.1 | Edit | Help

PUTATIVE cis-elements - HEXAMERS

YOUR_ROW	YOUR_IDENTIFIER	LOCI	MSU_MODEL_LOCUS	AAAAAA	AAAAAC	AAAAAG	CTCCTC	TCTCCT	TACATG
			WITH_MOTIF	41791	31476	34286	30808	25712	17820
			TOTAL_LOCI	68982	68982	68982	68982	68982	68982
			%_IN_GENOME	61	46	50	45	37	26
			LOCI_MATCHED	28	18	15	22	21	14
			LOCI_IN_LIST	33	33	33	33	33	33
			%_IN_LIST	84	54	45	66	63	42
1	LOC_Os08g14400.1	1	LOC_Os08g14400.1					1	
2	LOC_Os03g44810.1	1	LOC_Os03g44810.1	3		1			
3	LOC_Os04g34970.1	1	LOC_Os04g34970.1	10	1		1	3	1
4	LOC_Os08g43200.1	1	LOC_Os08g43200.1	3	1		1	1	
5	LOC_Os02g32590.1	1	LOC_Os02g32590.1	2		1	3		
6	LOC_Os09g29130.1	1	LOC_Os09g29130.1	7		1			1
7	LOC_Os02g52670.1	1	LOC_Os02g52670.1	4			2		

Figure 3. Microarray analysis workflow using Rice DB. (a) The summary output after a list of differentially expressed Affymetrix probe sets are entered into Rice DB. (b) The output after clicking on the expand spreadsheet icon in the 'Annotations' column or after typing 'Show annotations for…'. All columns show the annotations from the various sources (listed in Table 1). (c) The output after clicking the expand spreadsheet icon in the 'Expressed_in' column, or after typing 'Show expression profiles for…'. The normalized expression intensities across the 41 different developmental tissues are shown after log transforming and viewing using the custom heat map in MS Excel. (di) Output showing the genes containing the experimentally confirmed motifs after the expand icon was clicked in the 'Exp_shown_motifs' column. (dii) The output after the 'View hexamers for…' was entered for a shortlist of genes. This shows the numerical and percentage occurrence of the 4096 possible hexamers in the input gene list, as well as these occurrences in the genome.

rows are blank; however, the next two match LOC_Os08g14400.1 and LOC_Os03g44810.1 (Figure 3a). Given that LOC_Os03g44810.1 represents just one of thousands of genes annotated as 'expressed/hypothetical/unknown protein', functional annotations were specifically compiled for Rice DB from different sources (Annotations; Table 1). Thus, if a gene has an MSU putative function annotation of 'expressed protein', the RAP description, domain annotations (Interpro, Prosite, Pfam) and Genebins functional annotations (Goffard and Weiller, 2007) are also shown (Figure 3b). Additionally, the transcription factor databases (Gao *et al.*, 2006; Riano-Pachon *et al.*, 2007) and kinase databases (Dardick *et al.*, 2007) have also been incorporated into the Genebins annotations in Rice DB, and thus a wider net of annotation information can now be found (Figure 3b). In this way, it was revealed that despite the annotation of 'expressed protein', LOC_Os03g44810.1 encodes a transcription factor with a zinc-finger and SPX domain (Figure 3b), which also supports the TAIR annotation for its Arabidopsis orthologue (Figure 2).

Furthermore, clicking on the small spreadsheet icon in the 'expressed_in' column shows the normalized expression levels across development (Figure 3c). Note that the MSU RGAP database also presents transcript profiles and co-expression data for rice genes present on microarrays, also allowing users to extract expression profiles in a similar way. For LOC_Os08g14400.1 and LOC_Os03g44810.1, the expression intensities were exported from Rice DB, logarithm-transformed in Microsoft EXCEL and false-coloured as a heat map within Microsoft EXCEL, visualising the high expression in seed/germination (Figure 3c). Interestingly, the expression annotations for the Arabidopsis orthologues to these also show seed-specific expression (Figure 2). It is, however, important to point out that although this does occur for several genes, parallel expression between these species must be interpreted with caution, given the divergence between these two model species. In this way, numerous studies have used transcript data to gain an expression 'context' of a given set of genes (Howell *et al.*, 2009; Huang *et al.*, 2009; Narsai *et al.*, 2009, 2011; Taylor *et al.*, 2010), and Rice DB can also now facilitate this with ease.

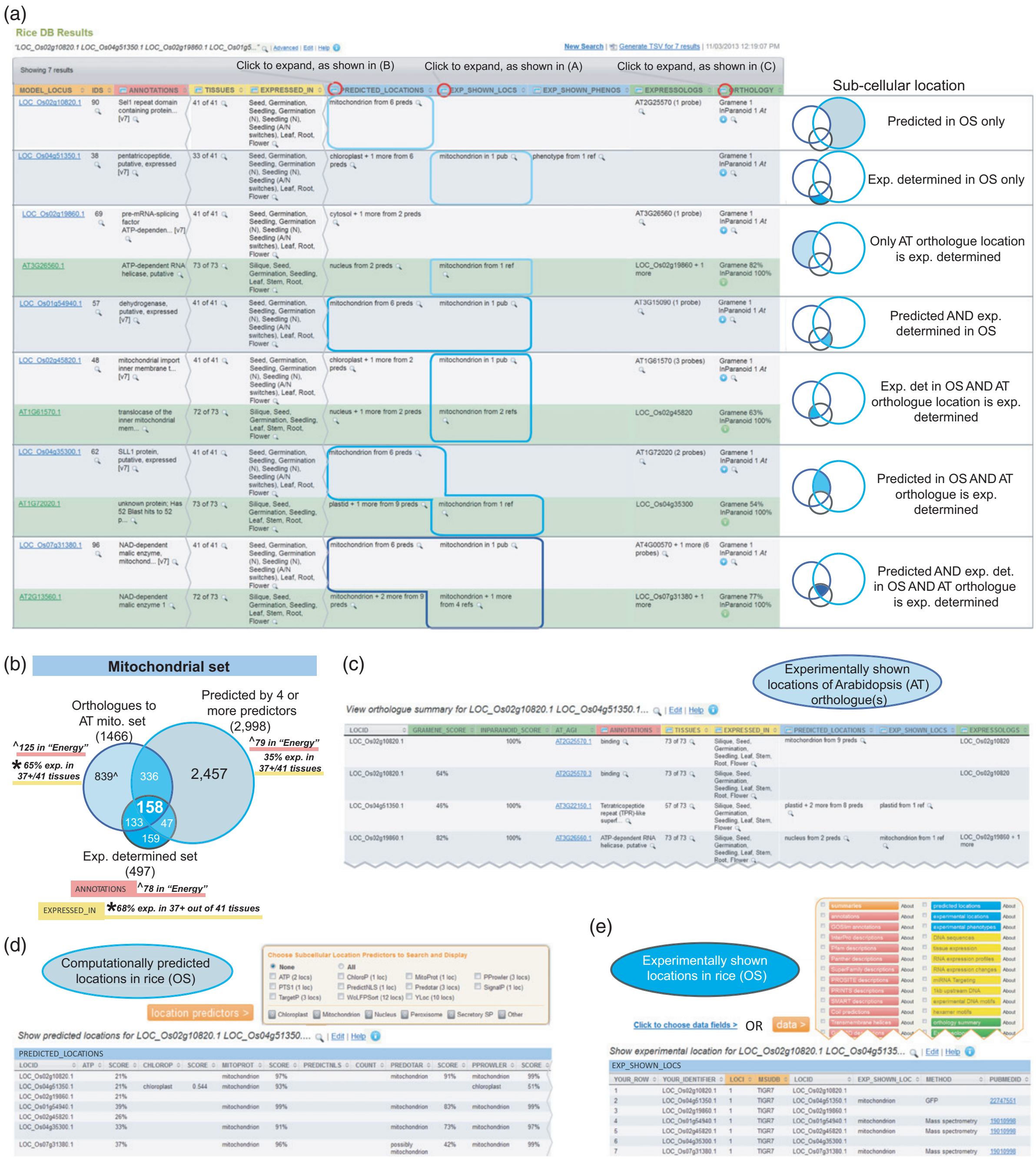
Following heat map/cluster generation in this way (Figure 3c), or even after identifying lists of co-expressed genes using other rice databases (e.g. ATTEDII (Obayashi *et al.*, 2011), Rice DB can be used to search for potential elements of co-regulation. It is possible to search the 1 kb upstream regions of a set of rice genes, both for the occurrence of experimentally identified (Bulow *et al.*, 2009; Yilmaz *et al.*, 2011) and putative motifs (Figure 3d). Figure 3(di) shows the output when experimentally demonstrated motifs are searched in the genes encoding transcription factors that showed highest expression in seed/germination. After sorting by EXP_SHOWN_MOTIF, sets of genes containing that motif, the motif sequence and source are shown (Figure 3di). Alternatively, a search such as 'show hexamers for [insert identifiers]' produces a table showing the occurrence of all 4096 possible hexamers in the upstream regions of these genes (Figure 3dii). Using percentage and occurrence values, the number of sequences containing a particular motif in a given subset (i.e. gene list) can be compared with the genome to reveal over-represented putative hexamers (Figure 3 dii). This function in Rice DB is comparable with the motif analysis tool available in TAIR (2005), where it is possible to simply search the 1 kb upstream regions of the Arabidopsis orthologous genes for hexamer occurrences in a subset compared to the genome. Thus, a common analysis workflow can be supported in Rice DB, from a differentially expressed set of genes to revealing putative regulatory elements of co-expression (Figure 3). Specifically, it was shown that LOC_Os08g14400.1 and LOC_Os03g44810.1 encode transcription factors (Figure 3b), which show seed/germination-specific expression (Figure 3c) and contain common RAV1–A, SORLIP3, RY-repeat and NAC motifs in their promoters (Figure 3di).

Subcellular localization of rice proteins: the rice mitochondrial proteome

One of the most important pieces of information required to define the function of a protein is its subcellular location. Isoforms of proteins can be located in different subcellular locations, and despite identical/conserved enzymatic activities, the functions can differ because of subcellular location. Examples include a variety of proteins located in mitochondria and plastids, which catalyse various metabolic steps in energy metabolism. Furthermore, some proteins can have multiple locations, termed dual targeting, and again, it is essential to know this in order to gain insight into function (Lu *et al.*, 2007).

In Arabidopsis, extensive knowledge of subcellular location has provided useful insight into specific protein functions. The SUBA database shows collated subcellular location information, compiling the outputs from multiple computational location predictors (analysing all protein sequences in the Arabidopsis genome) as well as linking to the collection of publications showing experimental evidence of localization, such as MS/MS and green fluorescence protein (GFP) analysis, which is manually curated (Heazlewood *et al.*, 2005). To date, no equivalent resource has been generated for rice, with a number of studies in rice only relying on the output from one or more predictors to gain insight into location (Lemberg and Freeman, 2007; Soanes *et al.*, 2012). Thus, in Rice DB, we not only incorporated the same two lines of evidence (computational prediction and experimentally determined localization), but also incorporated the subcellular location of orthologous proteins in Arabidopsis. Using Rice DB, a putative rice mitochondrial proteome is presented, comparing the three available methods for defining location, using specific examples (Figure 4a–d).

Firstly, all protein sequences encoded in the rice genome were analysed for subcellular location using 11 different computational predictors (listed in Table 1). Note that once a list of gene identifiers or the subcellular location is entered into the search box, and the predictors are selected (as in Figure 4d), the output shows the identifiers that matched the search criteria, including subcellular location and percentage/score for each predictor used (Figure 4d). For example, when 'mitochondria' is searched in Rice DB, 32 487 rows are returned and 23 501 of these are predicted by one or more predictors to be in the 'mitochondrion'. It is estimated that 1000–2000 proteins are present in mitochondria (Loreti *et al.*, 2003; Meisinger *et al.*, 2008). Thus, given the advantage of having a variety of predictor choices in Rice DB (Figure 4d), only proteins predicted to be mitochondrial by four or more predictors were included for comparison (2998; Figure 4b). For example, LOC_Os02g10820.1 was predicted to be mitochondrial by six predictors (predicted locations column; Figure 4a).


(a)
Click to expand, as shown in (B)
Click to expand, as shown in (A)
Click to expand, as shown in (C)
(b)
Mitochondrial set
Orthologues to AT mito. set (1466)
Predicted by 4 or more predictors (2,998)
^125 in "Energy"
* 65% exp. in 37+/41 tissues
^79 in "Energy"
35% exp. in 37+/41 tissues
839^
336
2,457
158
133
47
159
Exp. determined set (497)
ANNOTATIONS
^78 in "Energy"
EXPRESSED_IN
*68% exp. in 37+ out of 41 tissues
(c)
Experimentally shown locations of Arabidopsis (AT) orthologue(s)
(d)
Computationally predicted locations in rice (OS)
location predictors >
(e)
Experimentally shown locations in rice (OS)
Click to choose data fields >
OR
data >

Rice DB Results

"LOC_Os02g10820.1 LOC_Os04g51350.1 LOC_Os02g19860.1 LOC_Os01g5..." | Advanced | Edit | Help

New Search | Generate TSV for 7 results | 11/03/2013 12:19:07 PM

Showing 7 results

MODEL_LOCUS	IDS	ANNOTATIONS	TISSUES	EXPRESSED_IN	PREDICTED_LOCATIONS	EXP_SHOWN_LOCS	EXP_SHOWN_PHENOS	EXPRESSOLOGS	ORTHOLOGY	Sub-cellular location
LOC_Os02g10820.1	90	Sel1 repeat domain containing protein... [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	mitochondrion from 6 preds			AT2G25570 (1 probe)	Gramene 1 InParanoid 1 *At*	Predicted in OS only
LOC_Os04g51350.1	38	pentatricopeptide, putative, expressed [v7]	33 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	chloroplast + 1 more from 6 preds	mitochondrion in 1 pub	phenotype from 1 ref		Gramene 1 InParanoid 1 *At*	Exp. determined in OS only
LOC_Os02g19860.1	69	pre-mRNA-splicing factor ATP-dependen... [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	cytosol + 1 more from 2 preds			AT3G26560 (1 probe)	Gramene 1 InParanoid 1 *At*	Only AT orthologue location is exp. determined
AT3G26560.1		ATP-dependent RNA helicase, putative	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	nucleus from 2 preds	mitochondrion from 1 ref		LOC_Os02g19860 + 1 more	Gramene 82% InParanoid 100%	
LOC_Os01g54940.1	57	dehydrogenase, putative, expressed [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	mitochondrion from 6 preds	mitochondrion in 1 pub		AT3G15090 (1 probe)	Gramene 1 InParanoid 1 *At*	Predicted AND exp. determined in OS
LOC_Os02g45820.1	48	mitochondrial import inner membrane t... [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	chloroplast + 1 more from 2 preds	mitochondrion in 1 pub		AT1G61570 (3 probes)	Gramene 1 InParanoid 1 *At*	Exp. det in OS AND AT orthologue location is exp. determined
AT1G61570.1		translocase of the inner mitochondrial mem...	72 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	nucleus + 1 more from 2 preds	mitochondrion from 2 refs		LOC_Os02g45820	Gramene 63% InParanoid 100%	
LOC_Os04g35300.1	62	SLL1 protein, putative, expressed [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	mitochondrion from 6 preds			AT1G72020 (2 probes)	Gramene 1 InParanoid 1 *At*	Predicted in OS AND AT orthologue is exp. determined
AT1G72020.1		unknown protein; Has 52 Blast hits to 52 p...	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	plastid + 1 more from 9 preds	mitochondrion from 1 ref		LOC_Os04g35300	Gramene 54% InParanoid 100%	
LOC_Os07g31380.1	96	NAD-dependent malic enzyme, mitochond... [v7]	41 of 41	Seed, Germination, Seedling, Germination (N), Seedling (N), Seedling (A/N switches), Leaf, Root, Flower	mitochondrion from 6 preds	mitochondrion in 1 pub		AT4G00570 + 1 more (6 probes)	Gramene 1 InParanoid 1 *At*	Predicted AND exp. det. in OS AND AT orthologue is exp. determined
AT2G13560.1		NAD-dependent malic enzyme 1	72 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	mitochondrion + 2 more from 9 preds	mitochondrion + 1 more from 4 refs		LOC_Os07g31380 + 1 more	Gramene 77% InParanoid 100%	

View orthologue summary for LOC_Os02g10820.1 LOC_Os04g51350.1... | Edit | Help

LOCID	GRAMENE_SCORE	INPARANOID_SCORE	AT_AGI	ANNOTATIONS	TISSUES	EXPRESSED_IN	PREDICTED_LOCATIONS	EXP_SHOWN_LOCS	EXPRESSOLOGS
LOC_Os02g10820.1		100%	AT2G25570.1	binding	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	mitochondrion from 9 preds		LOC_Os02g10820
LOC_Os02g10820.1	64%		AT2G25570.3	binding	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower			LOC_Os02g10820
LOC_Os04g51350.1	46%	100%	AT3G22150.1	Tetratricopeptide repeat (TPR)-like superf...	57 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Flower	plastid + 2 more from 8 preds	plastid from 1 ref	
LOC_Os02g19860.1	82%	100%	AT3G26560.1	ATP-dependent RNA helicase, putative	73 of 73	Silique, Seed, Germination, Seedling, Leaf, Stem, Root, Flower	nucleus from 2 preds	mitochondrion from 1 ref	LOC_Os02g19860 + 1 more

Choose Subcellular Location Predictors to Search and Display

● None ○ All
□ ATP (2 locs) □ ChloroP (1 loc) □ MitoProt (1 loc) □ PProwler (3 locs)
□ PTS1 (1 loc) □ PredictNLS (1 loc) □ Predotar (3 locs) □ SignalP (1 loc)
□ TargetP (3 locs) □ WoLFPSort (12 locs) □ YLoc (10 locs)
□ Chloroplast □ Mitochondrion □ Nucleus □ Peroxisome □ Secretory SP □ Other

Show predicted locations for LOC_Os02g10820.1 LOC_Os04g51350.... | Edit | Help

PREDICTED_LOCATIONS

LOCID	ATP	SCORE	CHLOROP	SCORE	MITOPROT	SCORE	PREDICTNLS	COUNT	PREDOTAR	SCORE	PPROWLER	SCORE
LOC_Os02g10820.1		21%			mitochondrion	97%			mitochondrion	91%	mitochondrion	99%
LOC_Os04g51350.1		21%	chloroplast	0.544	mitochondrion	93%					chloroplast	51%
LOC_Os02g19860.1		21%										
LOC_Os01g54940.1		39%			mitochondrion	99%			mitochondrion	83%	mitochondrion	99%
LOC_Os02g45820.1		26%										
LOC_Os04g35300.1		33%			mitochondrion	91%			mitochondrion	73%	mitochondrion	97%
LOC_Os07g31380.1		37%			mitochondrion	96%			possibly mitochondrion	42%	mitochondrion	99%

□ summaries	About	□ predicted locations	About
□ annotations	About	□ experimental locations	About
□ GOSlim annotations	About	□ experimental phenotypes	About
□ InterPro descriptions	About	□ DNA sequences	About
□ Pfam descriptions	About	□ tissue expression	About
□ Panther descriptions	About	□ RNA expression profiles	About
□ SuperFamily descriptions	About	□ RNA expression changes	About
□ PROSITE descriptions	About	□ miRNA Targeting	About
□ PRINTS descriptions	About	□ 1kb upstream DNA	About
□ SMART descriptions	About	□ experimental DNA motifs	About
□ Coil predictions	About	□ hexamer motifs	About
□ Transmembrane helices	About	□ orthology summary	About

Show experimental location for LOC_Os02g10820.1 LOC_Os04g5135... | Edit | Help

EXP_SHOWN_LOCS

YOUR_ROW	YOUR_IDENTIFIER	LOCI	MSUDB	LOCID	EXP_SHOWN_LOC	METHOD	PUBMEDID
1	LOC_Os02g10820.1	1	TIGR7	LOC_Os02g10820.1			
2	LOC_Os04g51350.1	1	TIGR7	LOC_Os04g51350.1	mitochondrion	GFP	22747551
3	LOC_Os02g19860.1	1	TIGR7	LOC_Os02g19860.1			
4	LOC_Os01g54940.1	1	TIGR7	LOC_Os01g54940.1	mitochondrion	Mass spectrometry	19010998
5	LOC_Os02g45820.1	1	TIGR7	LOC_Os02g45820.1	mitochondrion	Mass spectrometry	19010998
6	LOC_Os04g35300.1	1	TIGR7	LOC_Os04g35300.1			
7	LOC_Os07g31380.1	1	TIGR7	LOC_Os07g31380.1	mitochondrion	Mass spectrometry	19010998

Figure 4. Subcellular location of rice proteins. (a) Seven genes are shown, representing combinations of the three ways that Rice DB can give insight into subcellular location: i.e. computational prediction ('Predicted in rice'); experimentally determined subcellular location of orthologous proteins in Arabidopsis ('Exp. shown in Arabidopsis'); and experimentally determined subcellular location of rice proteins ('Exp. shown in rice'). (b) Overlapping numbers of proteins identified as mitochondrial on the basis of these three approaches: i.e. (i) orthologues to AT mito. set; (ii) the Rice mito. set; and iiii) computational prediction based on four or more predictors. All three total sets, as well as the exclusive set of 839 proteins determined by orthology alone, were significantly enriched in the 'Energy' Genebins category ($P < 0.01$, indicated by ^), and the number of these in each set is indicated in brackets. Gene expression patterns for the genome (defined as all genes on the Affymetrix Rice genome microarray) and each of the three gene sets were examined. For each set, the percentage of genes expressed in none of the microarrays, between one and 36 of the tissues/stages and >90% of all tissues/developmental stages (i.e. more than 37 out of the 41 different developmental tissues/stages) is indicated. *Gene sets enriched in these proportions, compared with the genome. (c) The orthologue summary for the rice proteins identified as mitochondrial on the basis of orthology: orthologues to AT mito. set. (d) The pop-up search box where specific predictors can be selected. The outputs for each of these were presented as percentages (for most, where possible), and the expanded output of these is shown. (e) The pop-up window showing the possible data types that can be searched in Rice DB, when 'choose data type' is selected on the homepage. The example output is shown for experimentally determined locations. See references for predictors in the 'Data' pages in Rice DB.

Secondly, a multitude of publications were searched and lists of rice proteins with experimental evidence of location were carefully curated and compiled, including those with GFP analysis, MS/MS, immunogold labelling and immunodetection (usually encompassing organelle isolation and the use of specific antibodies). Collectively, over 500 publications relating to rice and the respective organelles (e.g. 'mitochondria') were searched for information about experimentally determined protein localization. In this way, a final list of 497 mitochondrial proteins was generated after compiling experimentally determined location information from different publications. For most publications, this also involved converting different protein identifiers into MSU identifiers. Also, all details of the experimental method used to determine localization had to be extracted and entered into Rice DB (e.g. 'Exp. determined set' for mitochondria, as shown in Figure 4b; see also Example 8.1 at the bottom of the Rice DB website). The proteins encoded by LOC_Os01g54940.1, LOC_Os02g45820.1 and LOC_Os07g31390.1 are examples of experimentally confirmed mitochondrial proteins (Figure 4a), and the expanded details of these are shown in Figure 4e. Alternatively, the advanced search window allows users to select the data to display: in this case, experimental location (Figure 4e). Additionally, for proteins with experimentally confirmed locations, Rice DB also shows a manually extracted brief description of the phenotype (if this is reported) when this gene is mutated (e.g. T–DNA insertion, EMS, TOS17 lines etc.), knocked-down (e.g. antisense, RNAi) or over-expressed (hyperlinked PubMed identifiers link directly to the publication). In this way, it is shown that the suppressed expression of several mitochondrial proteins results in developmentally impaired phenotypes: e.g. for OsARG (Figure 1; Ma *et al.*, 2012), DCW11 (Fujii and Toriyama, 2008), MIR (Ishimaru *et al.*, 2009) and a number of others.

Thirdly, although the 497 experimentally confirmed mitochondrial proteins represent a substantially extended list, it is still likely to represent less than half of all mitochondrial proteins in rice. Thus, orthologous Arabidopsis proteins are also shown, as a third method to gain insight into the possible location of rice proteins (Figure 4a–c). The orthologue summaries of three rice genes (from Figure 4a) show the strength of orthology and subcellular location of the Arabidopsis orthologues (Figure 4c). In this way, the 1265 Arabidopsis proteins considered to be mitochondrial were used (Law *et al.*, 2012), where ~72% of these have experimental evidence for mitochondrial localization (Heazlewood *et al.*, 2005), and the remaining 28% were predicted and considered to be mitochondrial on the basis of function (Law *et al.*, 2012). In order to do this without bias, no cut-offs or preferences were given to the method for determining orthology, which was defined based on sequence identity computed in Gramene (Youens-Clark *et al.*, 2011) and/or Inparanoid (Ostlund *et al.*, 2010) in Rice DB. Of the 1265 Arabidopsis mitochondrial proteins, 1164 have rice orthologues (making up the 1466 rice proteins, orthologous to the AT mito. set; Figure 4b). Additionally, there are numerous Arabidopsis membrane proteins with confirmed subcellular locations that cannot be accurately predicted by computational predictors, thus further supporting the usefulness of incorporating orthology. However, it is important to caution users against assuming direct conservation between Arabidopsis and rice for subcellular location, whereby despite orthology there can be divergence in the subcellular location of orthologous proteins, either because of technical differences in the experimental methods or because of real biological differences between these species (Xu *et al.*, 2013).

Thus, overlapping these lists for mitochondria (Figure 4b) allowed comparison of these methods. Note, although publications state that up to 60% of Arabidopsis and rice proteins with experimentally determined location are also predicted to be in that location (Heazlewood *et al.*, 2003; Huang *et al.*, 2009), this must not be mistaken to represent the accuracy of predictors. For example, although 302 of the 497 experimentally confirmed mitochondrial proteins (61%) were also predicted by MitoProt (Claros, 1995), this is only 2.6% (302 of 11 745) of all proteins predicted to be mitochondrial by MitoProt in rice, revealing a high false-positive rate for individual predictors (Figure S1). Similar overlaps were seen for PProwler and TargetP (<3%), and <5% for WoLFPSort and YLoc, whereas 7.6% was seen for Predotar (Figure S1; Hawkins and Boden, 2006; Emanuelsson *et al.*, 2007; Horton *et al.*, 2007; Briesemeister *et al.*, 2010). Thus using the four predictors combined, the range and power of these resulted in a 6.8% overlap with the 497 experimentally confirmed mitochondrial proteins (205 out of 2298; Figure 4b). Whereas ~20% (291) of the 1466 rice proteins orthologous to Arabidopsis mitochondrial proteins (Law *et al.*, 2012) overlapped with the experimentally confirmed rice mitochondrial proteins, pointing towards orthology as one of the more accurate ways of determining subcellular location in rice (Figure 4b).

To examine these mitochondrial lists in light of function and expression, we first took advantage of the compiled functional annotations that can be used in Rice DB. A simple search for the 'Energy in Genebins', revealed 638 MSU identifiers that were matched by Genebins annotations (Goffard and Weiller, 2007) to encode energy-related functions, which makes up 1.8% of the 35 787 genes annotated by Genebins. Simple matching these to each of the three sets from Figure 4(b) revealed that each of these were significantly enriched in 'Energy' functions ($P < 0.001$) (^denotes significant over-representation, compared with 1.8% in the genome; Figure 4a). Given that mitochondria are essential organelles for energy production, the significant over-representation of energy functions was not surprising, and in fact provided independent evidence supporting the accuracy of these sets as representing mitochondrial proteins. Notably, the subset of mitochondrial proteins defined exclusively by orthology was also enriched in energy functions (839 proteins; Figure 4b), whereas this was not seen for the mitochondrial subset based exclusively on prediction (2457 proteins; Figure 4b).

Additionally, gene expression was examined within each of the subsets (from Figure 4b) using the in–house generated Rice DB expression annotations. For all genes with probe sets in the rice genome, only 24% are expressed in >90% of all tissues (i.e. 37+ out of the 41 possible tissues in rice; Figure 4b). In contrast, 68% of the 497 experimentally confirmed rice mitochondrial proteins are expressed in >90% of all tissues, compared with the 24% in the genome (*significant, $P < 0.001$; over-representation denoted with an asterisk; Figure 4b). Similarly, a significant enrichment was also revealed for the 1466 genes orthologous to the AT mito. set, where >65% of these genes were expressed in >90% of all tissues (Figure 4b). In contrast, no such significant enrichment was observed for the genes encoding proteins predicted to be mitochondrial (35% of 2998; Figure 4b).

Notably, when the Arabidopsis RNA expression annotations in Rice DB were also examined, 65% of the AT mito. set were also expressed in >90% (66+ out of the 73) of the tissues analysed, which is also a significant ($P < 0.001$) enrichment of these genes compared with the 38% of Arabidopsis genes that show this expression in the Arabidopsis genome (data not shown), which is comparable with the observed enrichment in rice (Figure 4b). Given that several mitochondrial functions are essential for viability and central metabolism, it is not unexpected that these genes are expressed in most (>90%) tissues throughout plant development, and the ability to examine annotations, expression and subcellular locations like this in parallel, using Rice DB, can then strengthen the knowledge of given rice genes/proteins, especially where little other information is known.

Maximising insight by linking annotation, expression, regulation, subcellular location and orthology

As demonstrated in Figures 1–4, it is extremely useful to have multilevel knowledge incorporating annotations, transcript and protein data for both Arabidopsis and rice in parallel in Rice DB, as this maximizes insight in a way that is not currently possible using the existing rice databases. Furthermore, in contrast to most other rice databases, Rice DB acts as a portal linking through to several data sources, and users can start from any data type and link to an array of information for their genes/proteins of interest (Figure 5a). For example, Figure 3 demonstrates how it is possible for researchers analysing microarray or RNA sequencing data to have an assisted workflow by using Rice DB, following a common order of analysis. Also, although it was not presented in Figure 3, there is also additional transcript-related data in Rice DB, including annotated miRNA targets (Jeong *et al.*, 2011; Figure 5a; Table 1). Furthermore, by having expression data in parallel with protein data in Rice DB, it is possible to gain insight into multiple orthologous proteins, possibly revealing which of the multiple homologues may be functional in specific tissues or developmental stages (Figure 5). However, the differences between Arabidopsis and rice must also be taken into account for these comparisons, from the significant difference in genome size to considerable biological differences, which have led to 45% of rice genes not having significant Arabidopsis orthologues (See Combined Orthology 'Data' page).

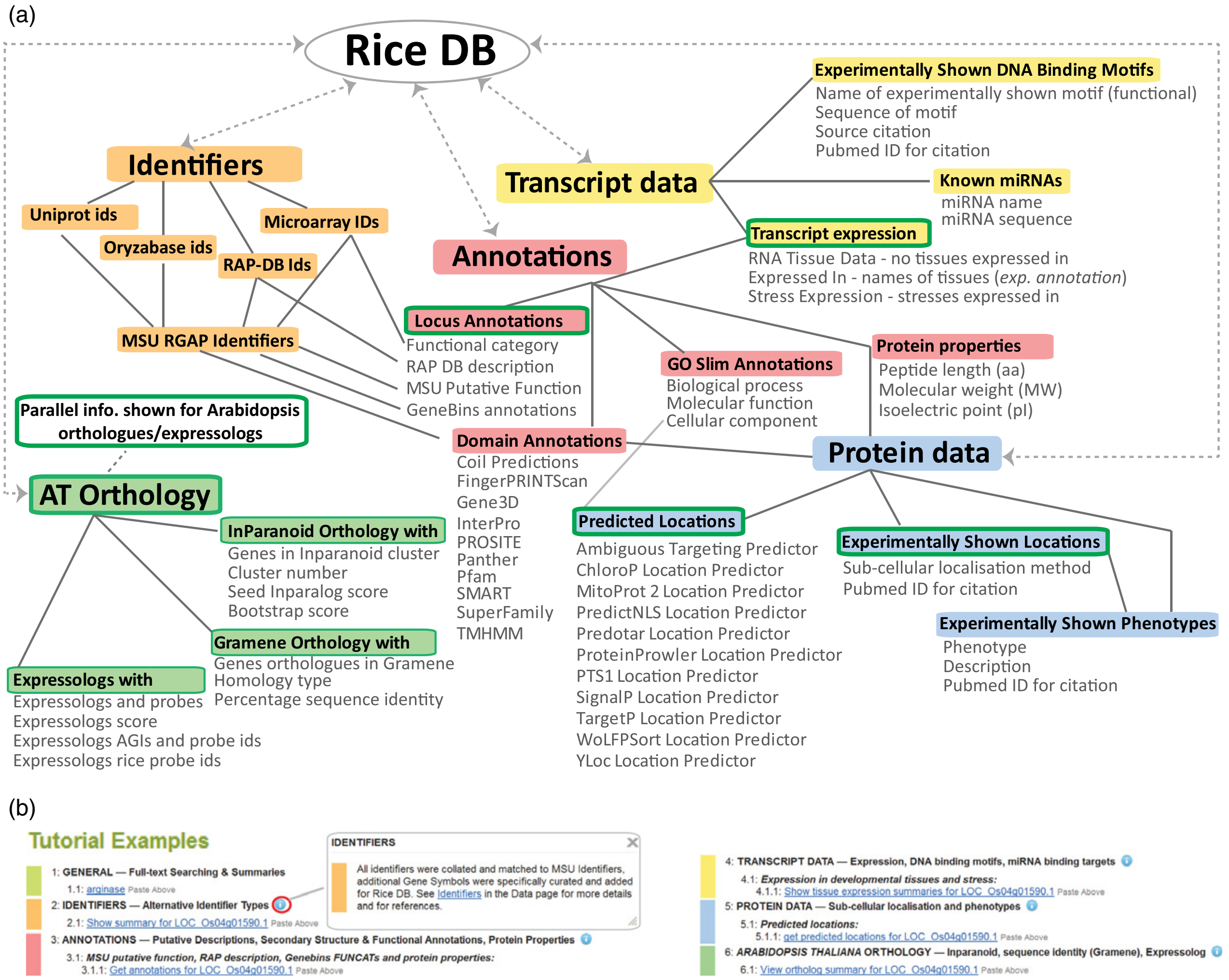
The flexibility of Rice DB also facilitates finding information in other workflows. For example, researchers examining specific transcription factor families may be interested in all genes containing a specific binding site: e.g. WRKY transcription factors and the W–box, TTGACC. Using Rice DB, it is possible to just enter that sequence into the search box and retrieve the list of 12 175 genes containing it in their promoters. Upon receiving this, it is possible to refine this (using the Refine tool) and identify any bias or over-representation(s). That is, it is possible to see if this set represents a co-expressed data set (using the transcript data), a set of co-localized proteins (using the subcellular location data) or a set of genes encoding proteins of a specific function (using the annotations) in Rice DB (Figure 5a; see examples in Figure 5b and Rice DB tutorial). Note, if the focus is co-expression, gene lists can also be easily exported and examined further in other databases, such as RiceFREND (Sato *et al.*, 2013a), Oryzaexpress (Hamada *et al.*, 2011) and RiceXPro (Sato *et al.*, 2013b), that specialize in detailed co-expression analysis.

The networked structure of Rice DB is also very useful to protein researchers. For example, protein properties can easily be retrieved using Rice DB, by simply entering 'Show protein properties for…' (Figure 5b; see Tutorial examples on Rice DB homepage). Following this, the peptide length, projected molecular weight and isoelectric point is shown for all proteins (Figure 5a). Thus, after receiving the protein properties for a list of proteins (e.g. those identified following mass spectrometry), it also possible to view and identify putative functional domains based on Gene3D, Interpro, Prosite and potential transmembrane helices based on TMHMM within Rice DB (Figure 5a; Table 1). Also, the computation and collation of predicted subcellular location(s) in Rice DB represents a resource not available anywhere else for rice (Figure 5a), and it is well known that subcellular location is extremely informative for defining protein function. The usefulness of combining data types, such as expression and subcellular location, is also demonstrated by its incorporation into the

BAR efp browser for Arabidopsis (Toufighi *et al.*, 2005). Lastly, the data/lists from Rice DB can also easily be exported for further searching in other resources such as the SALAD (Mihara *et al.*, 2010) or PRIN (Gu *et al.*, 2011) databases, which can reveal deeper insight into protein function by identifying conserved protein motifs or interactions.

Furthermore, the collation of phenotype information for proteins with known subcellular location represents another important resource in Rice DB, where rice phenotypes can simply be searched, revealing new trends and facilitating new hypotheses that would otherwise not have been apparent without Rice DB (Figure 5a). For example, a search for 'growth in experimental phenotypes' in Rice DB reveals 21 genes from 14 different publications, where genetic perturbation results in altered plant growth phenotypes. Viewing these closely revealed the two independent publications showing that mutating two different golgi-localized proteins results in growth alterations in plants (Li *et al.*, 2009; Zhang *et al.*, 2012). These were LOC_Os01g51430.1, which was annotated as 'green ripe-like, putative expressed protein', and LOC_Os12g36890.1, which was annotated as a 'cellulose synthase-like protein'. Thus, viewing these in parallel in Rice DB enables common threads to be identified. Likewise, clicking on the AGI of the closest Arabidopsis orthologue in Rice DB, quickly allows researchers to see if this has also been shown in Arabidopsis by opening the TAIR page, where links to publications relating to this gene are shown at the bottom of the page.

Also, once co-localized proteins are identified, Rice DB can also easily be used to gain insight into co-expression



1: GENERAL — Full-text Searching & Summaries
1.1: arginase Paste Above
2: IDENTIFIERS — Alternative Identifier Types
2.1: Show summary for LOC_Os04g01590.1 Paste Above
3: ANNOTATIONS — Putative Descriptions, Secondary Structure & Functional Annotations, Protein Properties
3.1: ***MSU putative function, RAP description, Genebins FUNCATs and protein properties:***
3.1.1: Get annotations for LOC_Os04g01590.1 Paste Above

IDENTIFIERS

All identifiers were collated and matched to MSU Identifiers, additional Gene Symbols were specifically curated and added for Rice DB. See Identifiers in the Data page for more details and for references.

4: TRANSCRIPT DATA — Expression, DNA binding motifs, miRNA binding targets
4.1: ***Expression in developmental tissues and stress:***
4.1.1: Show tissue expression summaries for LOC_Os04g01590.1 Paste Above
5: PROTEIN DATA — Sub-cellular localisation and phenotypes
5.1: ***Predicted locations:***
5.1.1: get predicted locations for LOC_Os04g01590.1 Paste Above
6: *ARABIDOPSIS THALIANA* ORTHOLOGY — Inparanoid, sequence identity (Gramene), Expressolog
6.1: View ortholog summary for LOC_Os04g01590.1 Paste Above

Figure 5. Inter-connections within Rice DB: *Oryza* information portal. (a) Rice DB creates a network for rice that connects identifiers, annotations, transcript data and protein data, and links these with information for orthologous genes in Arabidopsis. Data subtypes are shown below each heading. By connecting these data types for rice, it is possible to follow these connections and gain insight into function, including for rice genes with very little or no functional information. (b) Tutorial examples, as shown below the search box in Rice DB. These can be used as templates to use the functions in Rice DB. Note that only single examples are shown per data type (the full list is shown below the search box in Rice DB).

and co-regulation, without needing to manually translate identifiers or reformat searches for disparate specific database resources. In fact just having the different data types linked, such as the functional and expression annotations for both species in parallel, can yield insight, even for rice genes without detailed functional annotations. For example, At3g62790 is annotated as an NADH-ubiquinone oxidoreductase-related protein, and is expressed in all tissues, whereas its rice orthologue LOC_Os08g44250.1 is annotated as fiber protein Fb14, and is also expressed in all tissues. However, the annotation for At3g62790 and the knowledge that it is mitochondrial in Arabidopsis helps to provide greater insight into what may be the function of the protein encoded by the rice orthologue LOC_Os0844250.1. Thus, by using Rice DB, it possible to gain detailed insight into potential organellar proteins, which would not have been otherwise identified.

CONCLUSIONS

We have demonstrated a network of functional data relating to rice. Furthermore, in the style of search engines such as Google rice is now 'searchable', with little effort, and the retrieval of fundamental connections can now become routine and commonplace. Researchers can quickly gain access to rice knowledge by entering arbitrary identifiers, annotation keywords or even promoter motifs to immediately reveal relevant rice knowledge from previously unlinked data. In the presented examples, we have shown ways of exploring putative organellar proteomes through the use of localization information, expression and orthology for maximum insight into function. This is particularly important for genes and proteins with very little functional information, for which Rice DB can now reveal putative function annotations from a variety of sources (some improved upon and curated in-house), expression annotations (normalized and generated in-house), predicted subcellular localization (pre-computed in-house), experimentally determined location (collated and curated in-house) as well as the phenotypes (if any, upon mutation, knock-down, or overexpression) for proteins with experimentally confirmed subcellular locations (curated and collated in-house), and simply link to the Arabidopsis orthologues(s), where subcellular location and functional knowledge may already be known. Thus, Rice DB presents a simple, centralized data resource that can be used to gain maximum insight into rice gene/protein functions.

EXPERIMENTAL PROCEDURES

Database design

The Rice DB software and website have been developed for the JAVA runtime environment using the SCALA programming language and Lift web framework. The software's only operating system requirements are for a JAVA servlet runtime, disk storage and a network connection. See Appendix S1 for details.

Alternative identifiers

For the range of identifiers that can be searched in Rice DB, a variety of sources (outlined in Table 1: Alternative identifiers) were used, and a number were manually added. See Appendix S1 for details.

Annotations

We combined functional annotations with domain and structural annotations (outlined in Table 1) to maximize insight into putative function. A number of additional resources were also added and/or updated by manual collation and annotation. See Appendix S1 for details.

Transcript data: microarray analysis for expression annotations

To examine gene expression across development in rice (and Arabidopsis), a range of publically available microarrays were downloaded from the Gene Expression Omnibus or MIAME Array Express Databases (for each species), and these were analysed in-house using similar methods to previous studies (Howell *et al.*, 2009; Narsai *et al.*, 2011). See Appendix S1 for details.

Transcript data: microRNAs in rice

Known microRNA target genes, as identified in the Meyers lab Next-Gen Sequence Database (Jeong *et al.*, 2011), are also annotated to indicate known microRNA target genes in Rice DB.

Subcellular localization

To determine subcellular location, three main methods are presented, including: computational prediction, publications presenting experimental evidence and on the basis of orthology with Arabidopsis. Details of these (including thresholds used/cut-offs etc.) are detailed in Appendix S1. Note, although the mitochondrial proteome is shown as the example here, this information is also available for chloroplasts, peroxisomes, nucleus and various other organelles.

Compiling lists of genes with known phenotypes

For proteins with confirmed localizations based on experimental methods, phenotype details were also extracted for all genes that showed a phenotype when expression is altered: for example, by mutation (e.g. T–DNA insertion, EMS, TOS17 lines etc.); knockdown (e.g. antisense, RNAi); or overexpression, from the relevant publication. For these, a simple phenotype is shown in Rice DB: e.g. 'developmental' phenotype. Thus, for confirmed organellar proteins, a documented phenotype can also be searched in Rice DB. A description of the phenotype, as described in the publication, is shown in the column called 'Exp. shown. Pheno', and the relevant hyperlinked PubMed identifier is also shown.

Data sources

To view all the data sources used in Rice DB, see Table 1 and the 'Data' page (on the left panel of the Rice DB homepage).

ACKNOWLEDGEMENTS

This work was supported by an Australian Research Council Centre of Excellence Grant CEO561495, the Western Australian

State Government Centres of Excellence scheme and Super Science grant (FS100100022). The authors declare that they have no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Subcellular location of rice proteins using individual predictors.

Appendix S1. Details of the database design, inclusions and analyses methods used in Rice DB.

REFERENCES

Agrawal, G.K., Bourguignon, J., Rolland, N. *et al.* (2010) Plant organelle proteomics: Collaborating for optimal cell function. *Mass Spectrom. Rev.* **30**, 772–853.

Attwood, T.K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P.B., Popov, I., Roma-Mateo, C., Theodosiou, A. and Mitchell, A.L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)*, **2012**, bas019.

Borevitz, J.O. and Ecker, J.R. (2004) Plant genomics: the third wave. *Annu. Rev. Genomics Hum. Genet.* **5**, 443–477.

Briesemeister, S., Rahnenfuhrer, J. and Kohlbacher, O. (2010) YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.* **38**, W497–502.

Bulow, L., Engelmann, S., Schindler, M. and Hehl, R. (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res.* **37**, D983–986.

Chory, J., Ecker, J.R., Briggs, S. *et al.* (2000) National Science Foundation-Sponsored Workshop Report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* **123**, 423–426.

Claros, M.G. (1995) MitoProt, a Macintosh application for studying mitochondrial proteins. *Comput. Appl. Biosci.* **11**, 441–447.

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R. and Markley, J.L. (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **26**, 162–164.

Dardick, C., Chen, J., Richter, T., Ouyang, S. and Ronald, P. (2007) The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiol.* **143**, 579–586.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971.

Fujii, S. and Toriyama, K. (2008) DCW11, down-regulated gene 11 in CW-type cytoplasmic male sterile rice, encoding mitochondrial protein phosphatase 2c is related to cytoplasmic male sterility. *Plant Cell Physiol.* **49**, 633–640.

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., Gu, X., Wei, L. and Luo, J. (2006) DRTF: a database of rice transcription factors. *Bioinformatics*, **22**, 1286–1287.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H. and Geisler, M. (2007) A predicted interactome for Arabidopsis. *Plant Physiol.* **145**, 317–329.

Gibbs, D.J., Lee, S.C., Isa, N.M. *et al.* (2011) Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants. *Nature*, **479**, 415–418.

Goffard, N. and Weiller, G. (2007) GeneBins: a database for classifying gene expression data, with application to plant genome arrays. *BMC Bioinformatics*, **8**, 87.

Gu, H., Zhu, P., Jiao, Y., Meng, Y. and Chen, M. (2011) PRIN: a predicted rice interactome network. *BMC Bioinformatics*, **12**, 161.

Hamada, K., Hongo, K., Suwabe, K. *et al.* (2011) OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. *Plant Cell Physiol.* **52**, 220–229.

Han, B., Xue, Y., Li, J., Deng, X.W. and Zhang, Q. (2007) Rice functional genomics research in China. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 1009–1021.

Hawkins, J. and Boden, M. (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.* **4**, 1–18.

Heazlewood, J.L., Howell, K.A., Whelan, J. and Millar, A.H. (2003) Towards an analysis of the rice mitochondrial proteome. *Plant Physiol.* **132**, 230–242.

Heazlewood, J.L., Tonti-Filippini, J., Verboom, R.E. and Millar, A.H. (2005) Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. *Plant Physiol.* **139**, 598–609.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35**, W585–587.

Howell, K.A., Narsai, R., Carroll, A., Ivanova, A., Lohse, M., Usadel, B., Millar, A.H. and Whelan, J. (2009) Mapping metabolic and transcript temporal switches during germination in rice highlights specific transcription factors and the role of RNA instability in the germination process. *Plant Physiol.* **149**, 961–980.

Huang, S., Taylor, N.L., Narsai, R., Eubel, H., Whelan, J. and Millar, A.H. (2009) Experimental analysis of the rice mitochondrial proteome, its biogenesis, and heterogeneity. *Plant Physiol.* **149**, 719–734.

Ishimaru, Y., Bashir, K., Fujimoto, M., An, G., Itai, R.N., Tsutsumi, N., Nakanishi, H. and Nishizawa, N.K. (2009) Rice-specific mitochondrial iron-regulated gene (MIR) plays an important role in iron homeostasis. *Mol. Plant* **2**, 1059–1066.

Jaiswal, P., Ni, J., Yap, I. *et al.* (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.* **34**, D717–723.

Jeong, D.H., Park, S., Zhai, J., Gurazada, S.G., De Paoli, E., Meyers, B.C. and Green, P.J. (2011) Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell*, **23**, 4185–4207.

Joshi, H.J., Hirsch-Hoffmann, M., Baerenfaller, K. *et al.* (2011) MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. *Plant Physiol.* **155**, 259–270.

Jung, K.H., An, G. and Ronald, P.C. (2008a) Towards a better bowl of rice: assigning function to tens of thousands of rice genes. *Nat. Rev. Genet.* **9**, 91–101.

Jung, K.H., Dardick, C., Bartley, L.E. *et al.* (2008b) Refinement of light-responsive transcript lists using rice oligonucleotide arrays: evaluation of gene-redundancy. *PLoS ONE* **3**, e3337.

Klodmann, J., Senkler, M., Rode, C. and Braun, H.P. (2011) Defining the protein complex proteome of plant mitochondria. *Plant Physiol.* **157**, 587–598.

Kopka, J., Schauer, N., Krueger, S. *et al.* (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 1635–1638.

Koroleva, O.A., Tomlinson, M.L., Leader, D., Shaw, P. and Doonan, J.H. (2005) High-throughput protein localization in Arabidopsis using Agrobacterium-mediated transient expression of GFP-ORF fusions. *Plant J.* **41**, 162–174.

Kurata, N. and Yamazaki, Y. (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol.* **140**, 12–17.

Law, S.R., Narsai, R., Taylor, N.L., Delannoy, E., Carrie, C., Giraud, E., Millar, A.H., Small, I. and Whelan, J. (2012) Nucleotide and RNA metabolism prime translational initiation in the earliest events of mitochondrial biogenesis during Arabidopsis germination. *Plant Physiol.* **158**, 1610–1627.

Lemberg, M.K. and Freeman, M. (2007) Functional and evolutionary implications of enhanced genomic analysis of rhomboid intramembrane proteases. *Genome Res.* **17**, 1634–1646.

Letunic, I., Doerks, T. and Bork, P. (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302–305.

Li, M., Xiong, G., Li, R., Cui, J., Tang, D., Zhang, B., Pauly, M., Cheng, Z. and Zhou, Y. (2009) Rice cellulose synthase-like D4 is essential for normal cell-wall biosynthesis and plant growth. *Plant J.* **60**, 1055–1069.

Long, T.A., Brady, S.M. and Benfey, P.N. (2008) Systems approaches to identifying gene regulatory networks in plants. *Annu. Rev. Cell Dev. Biol.* **24**, 81–103.

Loreti, E., Yamaguchi, J., Alpi, A. and Perata, P. (2003) Sugar modulation of alpha-amylase genes under anoxia. *Ann. Bot.* **91 Spec No**, 3–148.

Lu, P.L., Chen, N.Z., An, R., Su, Z., Qi, B.S., Ren, F., Chen, J. and Wang, X.C. (2007) A novel drought-inducible gene, ATAF1, encodes a NAC family protein that negatively regulates the expression of stress-responsive genes in Arabidopsis. *Plant Mol. Biol.* **63**, 289–305.

Ma, X., Cheng, Z., Qin, R. ***et al.*** (2012) OsARG encodes an arginase that plays critical roles in panicle development and grain production in rice. *Plant J.* **73**, 190–200.

Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* **32**, D235–239.

Meisinger, C., Sickmann, A. and Pfanner, N. (2008) The mitochondrial proteome: from inventory to function. *Cell*, **134**, 22–24.

Meyer, E.H., Taylor, N.L. and Millar, A.H. (2008) Resolving and identifying protein components of plant mitochondrial respiratory complexes using three dimensions of gel electrophoresis. *J. Proteome Res.* **7**, 786–794.

Mihara, M., Itoh, T. and Izawa, T. (2010) SALAD database: a motif-based database of protein annotations for plant comparative genomics. *Nucleic Acids Res.* **38**, D835–842.

Mitschke, J., Fuss, J., Blum, T., Hoglund, A., Reski, R., Kohlbacher, O. and Rensing, S.A. (2009) Prediction of dual protein targeting to plant organelles. *New Phytol.* **183**, 224–235.

Narsai, R., Howell, K.A., Carroll, A., Ivanova, A., Millar, A.H. and Whelan, J. (2009) Defining core metabolic and transcriptomic responses to oxygen availability in rice embryos and young seedlings. *Plant Physiol.* **151**, 306–322.

Narsai, R., Castleden, I. and Whelan, J. (2010) Common and distinct organ and stress responsive transcriptomic patterns in Oryza sativa and Arabidopsis thaliana. *BMC Plant Biol.* **10**, 262.

Narsai, R., Law, S.R., Carrie, C., Xu, L. and Whelan, J. (2011) In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in Arabidopsis. *Plant Physiol.* **157**, 1342–1362.

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. and Eisenhaber, F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.* **328**, 581–592.

Nordborg, M. and Weigel, D. (2008) Next-generation genetics in plants. *Nature*, **456**, 720–723.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H. and Spannagl, M. (2013) MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* **41**, D1144–1151.

Obayashi, T., Nishida, K., Kasahara, K. and Kinoshita, K. (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.* **52**, 213–219.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O. and Sonnhammer, E.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–203.

Ouyang, S., Zhu, W., Hamilton, J. ***et al.*** (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–887.

Park, C.J., Bart, R., Chern, M., Canlas, P.E., Bai, W. and Ronald, P.C. (2010) Overexpression of the endoplasmic reticulum chaperone BiP3 regulates XA21-mediated innate immunity in rice. *PLoS ONE* **5**, e9262.

Patel, R.V., Nahal, H.K., Breit, R. and Provart, N.J. (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *Plant J.* **71**, 1038–1050.

Punta, M., Coggill, P.C., Eberhardt, R.Y. ***et al.*** (2012) The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–301.

Rice Genome. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Riano-Pachon, D.M., Ruzicic, S., Dreyer, I. and Mueller-Roeber, B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, **8**, 42.

Sato, Y., Namiki, N., Takehisa, H. ***et al.*** (2013a) RiceFREND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.* **41**, D1214–1221.

Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., Namiki, N., Ikawa, H., Ohyanagi, H., Sugimoto, K., Antonio, B.A. and Nagamura, Y. (2013b) RiceXPro version 3.0: expanding the informatics resource for rice transcriptome. *Nucleic Acids Res.* **41**, D1206–1213.

Schroder, F., Lisso, J. and Mussig, C. (2011) EXORDIUM-LIKE1 promotes growth during low carbon availability in Arabidopsis. *Plant Physiol.* **156**, 1620–1630.

Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W.B., Flugge, U.I. and Kunze, R. (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.* **131**, 16–26.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* **38**, D161–166.

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

Soanes, D.M., Chakrabarti, A., Paszkiewicz, K.H., Dawe, A.L. and Talbot, N.J. (2012) Genome-wide transcriptional profiling of appressorium development by the rice blast fungus Magnaporthe oryzae. *PLoS Pathog.* **8**, e1002514.

Swarbreck, D., Wilks, C., Lamesch, P. ***et al.*** (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–1014.

TAIR Arabidopsis genome update page (2005). http://www.arabidopsis.org/info/agicomplete.js.

Tanaka, T., Antonio, B.A., Kikuchi, S. ***et al.*** (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–1033.

Tanz, S.K., Castleden, I., Hooper, C.M., Vacher, M., Small, I. and Millar, H.A. (2013) SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.* **41**, D1185–1191.

Taylor, N.L., Howell, K.A., Heazlewood, J.L., Tan, T.Y., Narsai, R., Huang, S., Whelan, J. and Millar, A.H. (2010) Analysis of the rice mitochondrial carrier family reveals anaerobic accumulation of a basic amino acid carrier involved in arginine metabolism during seed germination. *Plant Physiol.* **154**, 691–704.

Thomas, P.D., Kejariwal, A., Campbell, M.J. ***et al.*** (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31**, 334–341.

Toufighi, K., Brady, S.M., Austin, R., Ly, E. and Provart, N.J. (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *Plant J* **43**, 153–163.

Wells, D.M., Wilson, M.H. and Bennett, M.J. (2010) Feeling UPBEAT about growth: linking ROS gradients and cell proliferation. *Dev. Cell* **19**, 644–646.

Xu, J., Yang, J., Wu, Z. ***et al.*** (2013) Identification of a dual-targeted protein belonging to the mitochondrial carrier family that is required for early leaf development in rice. *Plant Physiol.* **161**, 2036–2048.

Yasuda, H., Hirose, S., Kawakatsu, T., Wakasa, Y. and Takaiwa, F. (2009) Overexpression of BiP has inhibitory effects on the accumulation of seed storage proteins in endosperm cells of rice. *Plant Cell Physiol.* **50**, 1532–1543.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L. and Grotewold, E. (2011) AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res.* **39**, D1118–1122.

Youens-Clark, K., Buckler, E., Casstevens, T. ***et al.*** (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* **39**, D1085–1094.

Zhang, Q., Li, J., Xue, Y., Han, B. and Deng, X.W. (2008) Rice 2020: a call for an international coordinated effort in rice functional genomics. *Mol. Plant* **1**, 715–719.

Zhang, W., Zhou, X. and Wen, C.K. (2012) Modulation of ethylene responses by OsRTH1 overexpression reveals the biological significance of ethylene in rice seedling growth and development. *J. Exp. Bot.* **63**, 4151–4164.