Finding The Subcellular Location Of Barley, Wheat, Rice And Maize Proteins: The Compendium Of Crop Proteins With Annotated Locations (cropPAL)

Running head: Finding Crop Proteins With Annotated Locations

Corresponding authors:

Dr C. M. Hooper and Prof A. Harvey Millar

ARC Centre of Excellence in Plant Energy Biology

The University of Western Australia

M316, 35 Stirling Highway

Crawley WA 6009

Australia

Tel: +61 8 6488 4411

Fax: : +61 8 6488 4401

Email: cornelia.hooper@uwa.edu.au; harvey.millar@uwa.edu.au

Subject areas:

proteins, enzymes and metabolism

genomics, systems biology and evolution

Tables and Figures:

b/w figures: -

Color figures: 3

Tables: 1

Supplementary materials: 1

Finding The Subcellular Location Of Barley, Wheat, Rice And Maize Proteins: The Compendium Of Crop Proteins With Annotated Locations (cropPAL)

Running head: Finding Crop Proteins With Annotated Locations

Cornelia M Hooper, Ian R Castleden, Nader Aryamanesh, Richard P Jacoby and A Harvey Millar

ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, WA 6009, Australia

Abbreviations

cropPAL	crop Proteins with Annotated Locations
DFCI	Dana Faber Cancer institute
GFP	Green Fluorescent Protein
GUI	Graphical User Interface
KEGG	Kyoto Encyclopedia of Genes and Genomes
MS	Mass spectrometry
MSU	Michigan State University
RAP	Rice Annotation Project
SUBA	The SUBcellular localization database for Arabidopsis proteins

Footnotes

The cropPAL data collection was registered in the following two repositories: Research Data Online (https://researchdataonline.research.uwa.edu.au/handle/123456789/2141) and Research Data Australia (http://dx.doi.org/10.4225/23/556e4e260e093).

Abstract

Barley, wheat, rice and maize provide the bulk of human nutrition and have extensive industrial use as agricultural products. The genomes of these crops each contain more than 40,000 genes encoding proteins, however, the major genome databases for these species lack annotation information of protein subcellular location for >80% of these gene products. We address this gap, by constructing the compendium of crop protein subcellular locations called crop Proteins with Annotated Locations (cropPAL).

Subcellular location is most commonly determined by fluorescent protein tagging of live cells or mass spectrometry detection in subcellular purifications, but can also be predicted from amino acid sequence or protein expression patterns. The cropPAL database collates 556 published studies, from >300 research institutes in >30 countries that have been previously published, as well as compiling 8 pre-computed subcellular predictions for all *Hordeum vulgare*, *Triticum aestivum*, *Oryza sativa* and *Zea mays* protein sequences. The data collection including meta data for proteins and published studies can be accessed through a search portal http://crop-PAL.org.

The subcellular localization information housed in cropPAL helps to depict plant cells as compartmentalized protein networks that can be investigated for improving crop yield and quality, and developing new biotechnological solutions to agricultural challenges.

Keywords: Proteomes, Subcellular localizations, Database, Crop, Cell Biology, Compartments

Introduction

Recent developments in high-throughput genome sequencing technologies, computing and database management have made the protein sequences available for a range of non-model plant species of economic importance (http://www.gramene.org/). The number of reference plant genomes has tripled in only three years (Monaco et al. 2014) including the improved coverage of the bread wheat genome that has tackled a number of problems when annotating polyploidy genome sets (Bolser et al. 2015). The fact that the majority of plant proteins have sequence or functional motif relatives in other species opens up opportunities to link information of well-studied model or crop plants to proteins predicted from new genome sequences in order to advance more rapidly (Otto et al. 2008). The amount of gene and genome duplication and gene loss throughout angiosperms in the last 200 million years has led to huge variation in genome size between even close relatives (Tang et al. 2008). This raises questions around protein specialization and sub-functionalization of related proteins in plant cells.

One form of specialization of proteins with similar functions is their residence in distinct subcellular compartments. Small differences in protein sequence can allow optimal function in different compartments that may differ in pH or other conditions (Scheibe et al. 2005). The cost of duplication of proteins and pathways can outweigh the energy investment in transport across membranes (Cheung et al. 2013; Wu et al. 2006). Knowing about such changes in protein location during plant evolution has become crucial when targeting plant product yields and determining the energy budget of cells.

For the model plant Arabidopsis, the newly formed web resource AraPort (Krishnakumar et al. 2015) offers a wide variety of data types, links and query options for gene and protein information that allow networks of plant proteins to be explored. Collation of subcellular localization of Arabidopsis proteins with sequence information and functional annotation was begun a decade ago with the SUBcellular localization database for Arabidopsis proteins (SUBA) (Heazlewood et al. 2005; Heazlewood et al. 2007). It has since grown into a gold-standard subcellular data collection for this model plant, with experimental evidence for 26% of the Arabidopsis reference proteome and the remaining predicted computationally (Hooper et al. 2014; Tanz et al. 2013).

Linking proteins to the model plant Arabidopsis has been useful to bridge knowledge gaps and increase data confidence for rice and maize proteomes (Huang et al. 2013; Narsai et al. 2013).

For crop plants several annotation databases including http://www.maizegdb.org/ (Schaeffer et al. 2011), http://rice.plantbiology.msu.edu/ (Ouyang et al. 2007) and http://www.shigen.nig.ac.jp/rice/oryzabase/ (Kurata and Yamazaki 2006) exist but so far only Rice DB (http://ricedb.plantenergy.uwa.edu.au/) provides subcellular protein location information. Building Rice DB highlighted the challenges of collating different data resources, the lack of standardized identifiers and the rapid changes in annotation of newly sequenced crop species (Narsai et al. 2013).

Currently, the most valuable data for subcellular location are direct experimental pieces of evidence, which are scattered across published articles and their supplemental data. Previously published data linked to no or obsolete protein accession entries is common in these published works, making it no surprise that a comprehensive data resource linking experimental localization data to plant protein sequences only currently exists for Arabidopsis.

Recent efforts to extend experimental data collations beyond Arabidopsis have been demonstrated for gene co-expression (Obayashi et al. 2009) by inclusion of data from soybean, rice and other plant species (Obayashi et al. 2014). A similar extension for subcellular localizations data has not yet been reported, but a significant body of experimental protein localization studies exists for barley (Endler et al. 2006; Ploscher et al. 2011), wheat (Kamal et al. 2012; Suliman et al. 2013), rice (Natera et al. 2008; Reiland et al. 2011) and maize (Huang et al. 2013; Ma et al. 2010; Majeran et al. 2012). These crops have also been the focus of large-scale genome sequencing projects (Chapman et al. 2015; International Barley Genome Sequencing et al. 2012; International Rice Genome Sequencing 2005; International Wheat Genome Sequencing 2014; Schnable et al. 2009).

By generating the compendium of crop Proteins with Annotated Locations (cropPAL), we have connected experimental and predicted subcellular localizations for four major crop species proteomes into one database. We have extracted data from >556 studies and pre-computed subcellular localizations from eight predictors for barley (*Hordeum vulgare*), wheat (*Triticum aestivum*), rice (*Oryza sativa*) and maize (*Zea mays*). Using the Ensembl Plants/Gramene identifier system and a semi-automated linking pipeline, experimental data in cropPAL is linked to the current genome annotations offering sustainable links of otherwise static data with increasingly obsolete identifiers. We have generated a new web interface at http://crop-pal.org

that provides a central access location for querying and comparing localization data in all four crops.

Results

Protein sequence sets

We obtained the protein reference sets for the crop species from Ensembl Plants, accessed through the Gramene browser http://www.gramene.org/ (Monaco et al. 2014). The reference proteomes were used as the primary key for predicting subcellular location and linking experimental localizations. Updated iterations of these proteomes are periodically released. Ensembl Plants versions 23 and 26 were obtained during the period over which cropPAL was built and the differences in protein annotation were assessed. Genome annotations for rice, maize and barley were stable with no or few changes between the Ensembl Plants releases. In contrast, the wheat proteome annotation is still in progress and 10% of the proteins were re-annotated within the last year. For cropPAL, the final reference proteomes derived from version 26 contained 62,311, 99,354, 42,132 and 63,235 non-redundant sequences for barley, wheat, rice and maize, respectively.

To keep cropPAL up to date with future genome annotation changes, we developed a dynamic, semi-automated pipeline for maintaining links between localization data and the latest proteome sequences over sequential updates. This entailed isolating changes in annotations, re-computing any changed protein sequences and re-linking predicted and experimental location data to new reference proteins. The linking process was kept as transparent possible and is described below.

Experimental data collation

Literature search and study retrieval. The PubMed library (http://www.ncbi.nlm.nih.gov/pubmed) was searched for publications containing subcellular green fluorescent protein (GFP) and mass spectrometry (MS) data using methodology and species keywords (Supplementary Table S1). Using this keyword screening, approximately 3,000 publications were retrieved as a literature resource and manually screened for subcellular information to create a pre-filtered resource of ~1,000 studies across all species. Those studies assessed to contain GFP localization data were downloaded and stored in full text using Endnote X5. From each study, GFP data, protein name,

any provided database entry identification by the authors, target protein sequence, primer sequences used for the GFP construct, other experimental parameters and subcellular localization data were extracted into a primary entry file. The final data from 466 GFP studies were loaded into the cropPAL database as input data (Supplementary Table S2). All MS data containing studies were manually read and matched peptide to protein identifier lists, types of identifier provided and subcellular locations of proteins were extracted from the manuscript text or supplementary data. In total, data from 95 studies were extracted and linked to the Ensembl Plants identifier system.

Location categories. The experimental localizations of proteins were extracted using vocabulary control for 10 subcellular localizations 'cytoskeleton', 'cytosol', 'endoplasmic reticulum', 'extracellular', 'Golgi', 'mitochondrion', 'nucleus', 'peroxisome', 'plasma membrane', 'plastid', 'vacuole'. For display purposes the locations 'endoplasmic reticulum', 'extracellular', 'Golgi', 'plasma membrane' and 'vacuole' were also annotated as 'secretory'. Locations described in the literature as 'endosome' were assigned to 'Golgi', 'cell plate' was assigned to 'cytosol' and 'cell wall' to 'extracellular'. This is compatible with location calls used by SUBA3 (http://suba3.plantenergy.uwa.edu.au/) as well as Rice DB (http://ricedb.plantenergy.uwa.edu.au/) offering easy comparison with these data sets.

Database identifier collation. We found that at least six distinct identifier systems were used to describe similar of identical protein sequences of barley, wheat, rice and maize proteomes (Supplementary Table S3). The most commonly used identifier types included UniProt, NCBI, Michigan State University (MSU) (Ouyang et al. 2007), Dana Faber Cancer institute (DFCI) (Antonescu et al. 2010), Rice Annotation Project (RAP) (Sakai et al. 2013) and Ensembl Plants (Monaco et al. 2014). Due to progression in the annotation of sequence databases, static experimental data reported in the literature become disconnected from meaningful protein sequences and lose usability with time. This can lead to loss of the public availability of valuable experimental data that could be useful to current research. We have linked references to obsolete sequence systems such as DFCI as well as cross-linked and BLAST-linked multiple identifier

systems to the current Ensembl Plants identifier system. In this way cropPAL offers a valuable pipeline and access to subcellular localization data sets that are now directly searchable.

Experimental localization data in cropPAL. Many studies reporting subcellular location based on GFP experiments provide protein names rather than identifiers, which may lead to ambiguous links to genome annotations. Few localization studies assessed here included identifiers that derived from the main annotation systems previously mentioned. Where identifiers were given by the authors, the proteins were cross-referenced to the Ensembl Plants identifier system (Supplementary Figure S4). However, most GFP data were linked using PCR primer sequences for construct generation as described in the method section. The primer pairs or recovered NCBI transcript sequences were matched against the Ensembl Plants cDNA sequences version 26 using BLAST and the closest match was retrieved. Experimental data that could not be linked using cross-referencing or BLAST were omitted from cropPAL. Using this strategy we were able to link 1,157 GFP localizations to 855 distinct crop proteins (Table 1).

Localization data from MS studies were mostly found in table format in supplementary documentation. The published MS data for the four crop species used peptide matching to six distinct identifier systems (UniProt, NCBI, DFCI, Ensembl Plants, MSU and RAP). Since published MS data rarely contains matched peptide sequences, these data were linked using cross-references or BLAST methods to match sequences associated with the identifiers provided by the authors. In order to link as many identifiers as possible, we chose a multi-linking approach depending on the information provided in the studies (Supplementary Figure S5). The Ensembl Plants data set offers cross-reference links to UniProt identifiers. For rice the MSU identifiers were directly convertible using the RAP converter tool (http://rapdb.dna.affrc.go.jp/). If MS data was presented using systems other than MSU or UniProt, we first retrieved any possible cross-links to MSU or UniProt. Where no cross-reference was found, the protein sequence was obtained from the respective database and matched (BLASTP) against the Ensembl Plants crop sequences. MS data for which no cross-links or blast-matches could be retrieved were omitted from cropPAL. From all retrieved MS studies, we were able to assign 16,763 localizations describing 10,692 distinct crop proteins (Table 1).

Combined, all experimental data in cropPAL comprises 17,920 localizations for 11,375 distinct proteins across all four crops. The coverage varies, with maize achieving the highest proteome coverage (9.5%) and barley the lowest (0.5%) from experimental localizations. Coverage of rice (8.1%) and maize in cropPAL version 1.2 is slightly higher than the initial coverage of the first collection housed in SUBA1 (7%) (Heazlewood et al. 2007). With increasing interest in subcellular studies in crop species and the development of GFP/MS technology we anticipate proteome coverage in crop species to rapidly increase towards the current proteome coverage for Arabidopsis housed in SUBA3 (26%) (Hooper et al. 2014).

Pre-computed prediction data integration

In all species, experimental localization data still only covers a limited portion of the predicted proteomes. In order to provide a comprehensive collection of localization data in cropPAL, we sourced outputs for every predicted protein from eight predictor algorithms. The Ensembl Plants proteome sequences for all four crop plants were used to query BaCeLo (Pierleoni et al. 2006), ChloroP 1.1 (Emanuelsson et al. 1999), iPSORT (Bannai et al. 2002), PProwler 1.2 (Hawkins and Boden 2006), Predotar v1.03 (Small et al. 2004), TargetP 1.1 (Emanuelsson et al. 2000), WoLF PSORT (Horton et al. 2007) and YLoc (Briesemeister et al. 2010). The subcellular location calls from each predictor for each protein stored in the cropPAL database comprise 2,153,828 predictions in total (Supplementary Table S6). The combined set of location predictions gapfills when no experimental data is available. All subcellular data (predictive and experimental), protein annotations, protein properties, sequence information, and geographical or affiliation information from experimental studies were connected in a relational database that can be queried by the user through a web interface (Figure 1).

The cropPAL search interface

Search interface. We have developed a publicly accessible web interface for the cropPAL data at http://crop-pal.org which gives access to all localization calls for crop proteins ('Subcellular Location') as well as amino acid sequence metadata ('Protein Properties'), metadata associated with the experimental studies reporting their location ('Affiliations') and links to homologous proteins within the four crop species and Arabidopsis ('Homology'). The cropPAL interface offers

straightforward ‘Quick Search’ functions for subcellular proteomes as well as more complex query-building options (Figure 2A). The user can search individual crop species or several species can be searched in series. The cropPAL home page contains a single window (‘Quick Search’) for rapid searching of text or protein identifiers. For more complex searches, users can take advantage of the query builder tabs in the bottom half of the home page (Figure 2A). The top menu shows links to information about the cropPAL experimental data (“cropPAL stats”), the origin of the integrated studies (“Localization World Map”) as well as a link to a tutorial (“Tutorial”) on how to use the query builder.

Subcellular Location queries. CropPAL allows searching for experimental and/or predictive data while specifying inclusion/exclusion of any location, single predictor, any prediction or choosing GFP, MS or both data types. For starting a query and adding multiple filters, the ‘+’ button can be used and the query will appear in the query window. The query window displays options to add more filters using ‘AND’ or ‘OR’ and appropriate brackets (Figure 2A).

Protein Properties queries. For adding filters regarding protein sequence and gene loci properties users can choose the “Protein Properties” tab. This contains several options to filter for protein descriptions, number of amino acids, molecular weight, isoelectric point, hydrophobicity (GRAVY) (Kyte and Doolittle 1982), location on a specific assembly or chromosome, or functional annotations including GO terms, interpro and Pfam domains.

Affiliations queries. Each experimental localization in cropPAL remains linked to the original publication and thus to the authors, their geographical location, year of publication, abstract and title as well as the PubMed reference of the study. These data can be searched or used as a filter, and can enhance data interpretation or enable the connection of researchers.

Homology queries. The “Homology” tab provides links between crop protein localization data and comparable localization data for Arabidopsis orthologs. The ortholog link between crops and Arabidopsis was generated using reciprocal BLASTX of crop protein sequence against the Arabidopsis sequence database and the best match by score, probability and length was

retained. We were able to generate best match links between Arabidopsis proteins and 60,777 barley, 96,901 wheat, 38,580 rice and 59,912 maize proteins. The alignment between the protein sequences is displayed in the factsheet for users to evaluate the BLASTX fit.

For the crop-to-crop comparison of sequences, we used the homology tree from the Ensembl Plants mart (see methods). The Ensembl Plants homology search allows retrieval of paralogs as well as orthologs of any species combination in a specific subcellular location. The user can choose a cut off for both methodologies using either BIT score (reciprocal BLASTX) or by sequence identity (Ensembl Plants homology tree).

cropPAL multi-output interface

Results tabs. The cropPAL interface opens a result tab for each species and each query submitted. The results are returned in table format that allow an overview of the returns based on protein identifier, description, predictions, experimental data, all homologs as assigned by Ensembl Plants as well as the best match in Arabidopsis (reciprocal BLASTX) with the SUBA consensus (SUBAcon) location and its description according to TAIR10 (Figure 2B). The default result view can be optimized to show other features including more detailed publication information, length, isoelectric point, GRAVY or molecular weight for each protein. For downstream usability, cropPAL results can be downloaded in csv format using the download button.

Factsheet for detailed protein information. For more detailed information the user can open the hyperlinked factsheet for each protein locus in the result view. The factsheet offers more detailed information of prediction calls made by individual predictors as well as more information about the localization studies (Figure 3). In the top left of the factsheet, the collated localization data from all predictors and experimental studies for the protein are summarized as the unweighted sum of location calls and a schematic visualization shows the most often “voted” location on a white-yellow-red scale.

In addition, the factsheet shows a variety of protein annotations including the current Ensembl Plants description, other protein aliases, sequence motives, features and annotations from other databases. Most of these are hyperlinked for rapid direct retrieval of more detailed information

about the associated feature. Besides the physical properties of the protein, the factsheet also offers the best match between the crop protein and Arabidopsis as a sequence alignment along with the Arabidopsis protein description and its SUBAcon location (Hooper et al. 2014).

Discussion

The compendium of cropPAL addresses a challenge in biological data management, by providing protein localization data linked to contemporary genome databases. Over the last decade, numerous consortia have produced genome sequence databases for the major crops barley, wheat, rice and maize. However, these databases provide very little information regarding protein subcellular localization in these important crop species. By building the cropPAL database that collates localization data derived from GFP and MS studies as well as computational prediction we offer this information. Building this database has involved retrieving several thousand experimental results from 556 distinct published reports, many of which were not linked to current databases due to obsolete or poorly-defined protein accessions. We have implemented a semi-automated pipeline to link these data to current and subsequent genebuild updates, ensuring that these valuable data can continue to be accessed in the future. The cropPAL data and interface (http://crop-pal.org) also offers a central access point for eight pre-computed subcellular location predictions coupled to this experimental evidence. By combing literature curation, relinking to current proteome annotations, computing predictor calls and homology integration we have generated an online aggregated resource on subcellular location of crop proteins. Combining data for a protein of interest offers users an overview of localization calls, recognizing that each predictor or experiment can harbor its own individual strengths and weaknesses (Hooper et al. 2014). The cropPAL search interface allows both simple and complex Boolean queries across all crop data sets. The search options can be used to build subcellular proteomes, compare performance of different localization methods and assess the location of user-defined sets of proteins. The collated subcellular localization information aims to aid research on plant cell compartmentalization in the future.

Methods

Reference proteomes and annotation resources

The version 23 and version 26 reference cDNA, protein sequences and annotation for barley (*Hordeum vulgares*), wheat (*Tritium aestivum*), rice (*Oryza sativa*) and maize (*Zea mays*) were downloaded from the Ensembl Plants server (http://plants.ensembl.org/info/website/ftp/index.html). The plant species homology database Ensembl "plant mart" version 26 was downloaded and locally installed. The database was cropped to tables containing only Arabidopsis, barley, wheat, rice and maize data. Sequence, annotation, subcellular localizations and subcellular consensus call SUBAcon for Arabidopsis proteins were sourced from SUBA3 (http://suba3.plantenergy.uwa.edu.au/). The DFCI gene indices for barley (version 9), maize (version 8, 14, 16, 17) and wheat (version11, 12) were obtained from the TGI database archives at the Dana Faber Institute and Harvard school of public health (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/). The data set contained DFCI indices, UniProt identifiers as well as the gene sequences. UniProt protein identification indices, cross-references and protein features for the four crop species were obtained from the UniProt server (http://www.UniProt.org/).

BLAST matching

Homology. In order to match an Arabidopsis ortholog to most crop proteins, we performed a reciprocal BLASTX between the Arabidopsis and either of the four crop sequences (bidirectional best hits). The top reciprocal match was obtained using the top score as a measure for fit in respect to fit length (Salichos and Rokas 2011; Wolf and Koonin 2012). The crop-to-crop homology tree version 26 was derived from Ensembl Plants (http://www.ensembl.org/info/genome/compara/homology_method.html) and is based on TreeBeST representing evolutionary relationships of gene families (Vilella et al. 2009).

Primer BLASTN. As described previously (Camacho et al. 2009), primer sequences were matched (BLASTN version 2.2.31, last updated May 2015) against cDNA sequences (Ensembl Plants version 26) using the settings recommended by NCBI (http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml). In short: word-size set at 7, "dust" filter off and e-value set at 1000. The top 10 sequences based on the best high-scoring segment pairs (HSP) scores were included and the identifier-intersect between the forward and reverse primers were collected.

Protein and transcript BLAST. In the absence of primer sequences and cross-references, the transcript sequences were matched to the cDNA sequences (Ensembl Plants version 26) using the settings recommended by NCBI for BLASTN (blast version 2.2.31) (Camacho et al. 2009). Protein sequences were matched against protein sequences (Ensembl Plants version 26) using the settings recommended by NCBI for BLASTP (blast version 2.2.31).

DFCI sequence blasts. The DFCI gene fragment sequences were translated and matched against the protein sequences (Ensembl Plants version 26) using the settings recommended by NCBI for BLASTX (blast version 2.2.31), as described previously (Camacho et al. 2009).

For all BLAST methods, only the hits with the highest match score were included in the cropPAL data set. Where multiple protein localizations in a single study led to one identical protein in Ensembl Plants proteome, the data were treated as a single localization.

Database structure and interface

CropPAL data are stored in a MySQL relational database, operating on a UNIX-based system. Extra relational semantics were added using the python SQLAlchemy library. The cropPAL search portal consists of a web-browser based GUI (Graphical User Interface) written in dynamic HTML that uses Asynchronous JavaScript + XML (AJAX) to interact with the cropPAL server.

The cropPAL search portal has been designed as a GUI for users without prior knowledge of SQL and allows the construction of complex queries in a point and click manner selecting from the query menu tabs. The GUI is accessible via http://crop-pal.org and is compatible with most current browsers. The keyword search window and query tabs are designed for ease of use. The tabs contain pre-formulated queries in full text with pull down menus and simple text boxes. Alternatively, the AND, OR and bracketing can be used to create complex Boolean queries.

Query results can be accessed via the results tabs in a tabular format with a default row format containing protein identifiers, localizations, homology matches and best Arabidopsis match functional annotation. Each column contains a pull down menu for customizing information shown. The results from a query may be downloaded as a tab limited file or Excel spreadsheet for further analysis. Each protein match in the results view is hyperlinked to a factsheet that opens a view to further information including sequence, hydropathy profile, protein size, coordinates of

gene and localization study description. Protein properties are displayed as hyperlinks that provide rapid access to related resources such as UniProt, AmiGO, EMBL, Ensembl Plants/ Gramene, KEGG, MSU, and others.

Funding

This project was supported by the Australian National Data Service (ANDS) through the National Collaborative Research Infrastructure Strategy Program, as well as through the ARC Centre of Excellence in Plant Energy Biology (CE140100008), AHM is supported as an ARC Future Fellow (FT110100242).

Disclosure

No conflict of interest declared.

Acknowledgments

The authors would like to acknowledge Katina Toufexis and Catherine Clark from the UWA Library for their advice and help with the data repository and generation of DOIs.

References

Antonescu, C., Antonescu, V., Sultana, R. and Quackenbush, J. (2010) Using the DFCI gene index databases for biological discovery. *Curr. Protoc. Bioinformatics* Chapter 1: Unit1 6 1-36.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. and Miyano, S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* 18: 298-305.

Bolser, D.M., Kerhornou, A., Walts, B. and Kersey, P. (2015) Triticeae resources in Ensembl Plants. *Plant Cell Physiol.* 56: e3.

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

Chapman, J.A., Mascher, M., Buluc, A., Barry, K., Georganas, E., Session, A., et al. (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* 16: 26.

Cheung, C.Y., Williams, T.C., Poolman, M.G., Fell, D.A., Ratcliffe, R.G. and Sweetlove, L.J. (2013) A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant J.*

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300: 1005-1016.

Emanuelsson, O., Nielsen, H. and von Heijne, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 8: 978-984.

Endler, A., Meyer, S., Schelbert, S., Schneider, T., Weschke, W., Peters, S.W., et al. (2006) Identification of a vacuolar sucrose transporter in barley and Arabidopsis mesophyll cells by a tonoplast proteomic approach. *Plant Physiol.* 141: 196-207.

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., 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.

Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I. and Millar, A.H. (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res.* 35: D213-218.

Hooper, C.M., Tanz, S.K., Castleden, I.R., Vacher, M.A., Small, I.D. and Millar, A.H. (2014) SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. *Bioinformatics* 30: 3356-3364.

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

Huang, M., Friso, G., Nishimura, K., Qu, X., Olinares, P.D., Majeran, W., et al. (2013) Construction of plastid reference proteomes for maize and Arabidopsis and evaluation of their orthologous relationships; the concept of orthoproteomics. *J. Proteome Res.* 12: 491-504.

International Barley Genome Sequencing, C., Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491: 711-716.

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

International Wheat Genome Sequencing, C. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 345: 1251788.

Kamal, A.H., Cho, K., Komatsu, S., Uozumi, N., Choi, J.S. and Woo, S.H. (2012) Towards an understanding of wheat chloroplasts: a methodical investigation of thylakoid proteome. *Mol. Biol. Rep.* 39: 5069-5083.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., Kim, M., et al. (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Res.* 43: D1003-1009.

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

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132.

Ma, W., Muthreich, N., Liao, C., Franz-Wachtel, M., Schutz, W., Zhang, F., et al. (2010) The mucilage proteome of maize (Zea mays L.) primary roots. *J. Proteome Res.* 9: 2968-2976.

Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., et al. (2012) Nucleoid-enriched proteomes in developing plastids and chloroplasts from maize leaves: a new conceptual framework for nucleoid functions. *Plant Physiol.* 158: 156-189.

Monaco, M.K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S., et al. (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.* 42: D1193-1199.

Narsai, R., Devenish, J., Castleden, I., Narsai, K., Xu, L., Shou, H., et al. (2013) Rice DB: an Oryza Information Portal linking annotation, subcellular location, function, expression, regulation, and evolutionary information for rice and Arabidopsis. *Plant J.* 76: 1057-1073.

Natera, S.H., Ford, K.L., Cassin, A.M., Patterson, J.H., Newbigin, E.J. and Bacic, A. (2008) Analysis of the Oryza sativa plasma membrane proteome using combined protein and peptide fractionation approaches in conjunction with mass spectrometry. *J. Proteome Res.* 7: 1159-1187.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.* 37: D987-991.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., et al. (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.* 55: e6.

Otto, T.D., Guimaraes, A.C., Degrave, W.M. and de Miranda, A.B. (2008) AnEnPi: identification and annotation of analogous enzymes. *BMC Bioinformatics* 9: 544.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35: D883-887.

Pierleoni, A., Martelli, P.L., Fariselli, P. and Casadio, R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* 22: e408-416.

Ploscher, M., Reisinger, V. and Eichacker, L.A. (2011) Proteomic comparison of etioplast and chloroplast protein complexes. *J. Proteomics* 74: 1256-1265.

Reiland, S., Grossmann, J., Baerenfaller, K., Gehrig, P., Nunes-Nesi, A., Fernie, A.R., et al. (2011) Integrated proteome and metabolite analysis of the de-etiolation process in plastids from rice (Oryza sativa L.). *Proteomics* 11: 1751-1763.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54: e6.

Salichos, L. and Rokas, A. (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS ONE* 6: e18755.

Schaeffer, M.L., Harper, L.C., Gardiner, J.M., Andorf, C.M., Campbell, D.A., Cannon, E.K., et al. (2011) MaizeGDB: curation and outreach go hand-in-hand. *Database* 2011: bar022.

Scheibe, R., Backhausen, J.E., Emmerlich, V. and Holtgrefe, S. (2005) Strategies to maintain redox homeostasis during photosynthesis under changing conditions. *J. Exp. Bot.* 56: 1481-1489.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112-1115.

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.

Suliman, M., Chateigner-Boutin, A.L., Francin-Allami, M., Partier, A., Bouchet, B., Salse, J., et al. (2013) Identification of glycosyltransferases involved in cell wall synthesis of wheat endosperm. *J. Proteomics* 78: 508-521.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science* 320: 486-488.

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.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19: 327-335.

Wolf, Y.I. and Koonin, E.V. (2012) A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* 4: 1286-1294.

Wu, S., Schalk, M., Clark, A., Miles, R.B., Coates, R. and Chappell, J. (2006) Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. *Nat. Biotechnol.* 24: 1441-1447.

Tables

Table 1. Experimental localization data in cropPAL

	SUBA3	**cropPAL**			
Species	*A. thaliana*	*H. vulgare*	*T. aestivum*	*O. sativa*	*Z. mays*
Number of localizations					
GFP	4107	92	160	691	214
MS	24142	436	2195	4353	5759
Total	28249	528	2355	5681	9356
Number of distinct proteins					
GFP	2644	53	113	527	162
MS	7893	282	1534	2673	5848
Total	9318	334	1635	3417	5989
Experimental coverage					
Proteome size[$]	*35388*	*62311*	*99354*	*42132*	*63235*
Proteome coverage	*26.3%*	*0.5%*	*1.6%*	*8.1%*	*9.5%*

[$] Numbers derived from http://www.gramene.org/ version 26. *GFP – green fluorescent protein;*

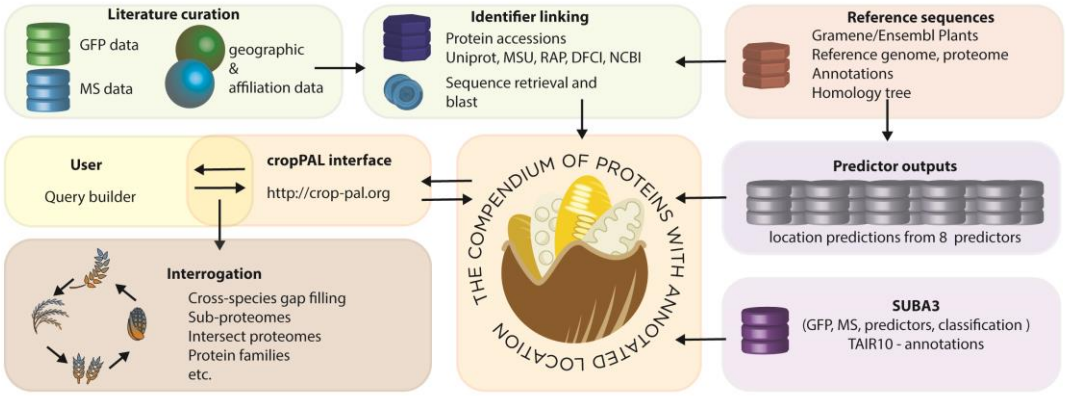
MS – mass spectrometry

Legends to figures

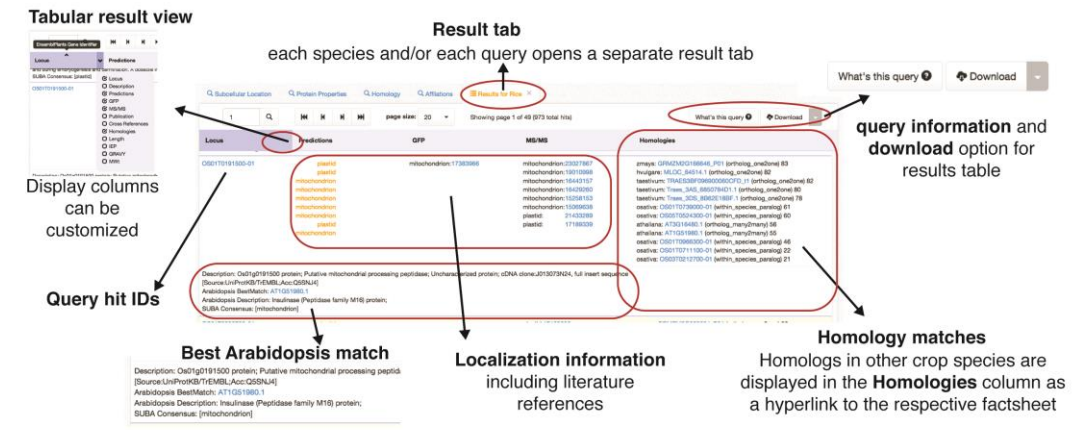
Fig. 1 Data curation, calculation and table relation in cropPAL. The green boxes (top left) highlight the experimental data curation, identifier linking and collation into the experimental data tables in cropPAL. The red, blue and purple boxes contain predicted and sourced localization and annotation data of the four crop plants and Arabidopsis for cropPAL. The structure of user interaction is schematically indicated in the yellow and brown boxes on the lower left. *GFP- Green Fluorescent Protein; MS – Mass Spectrometry; SUBA3 – SUBcellular localization database for Arabidopsis proteins version 3*

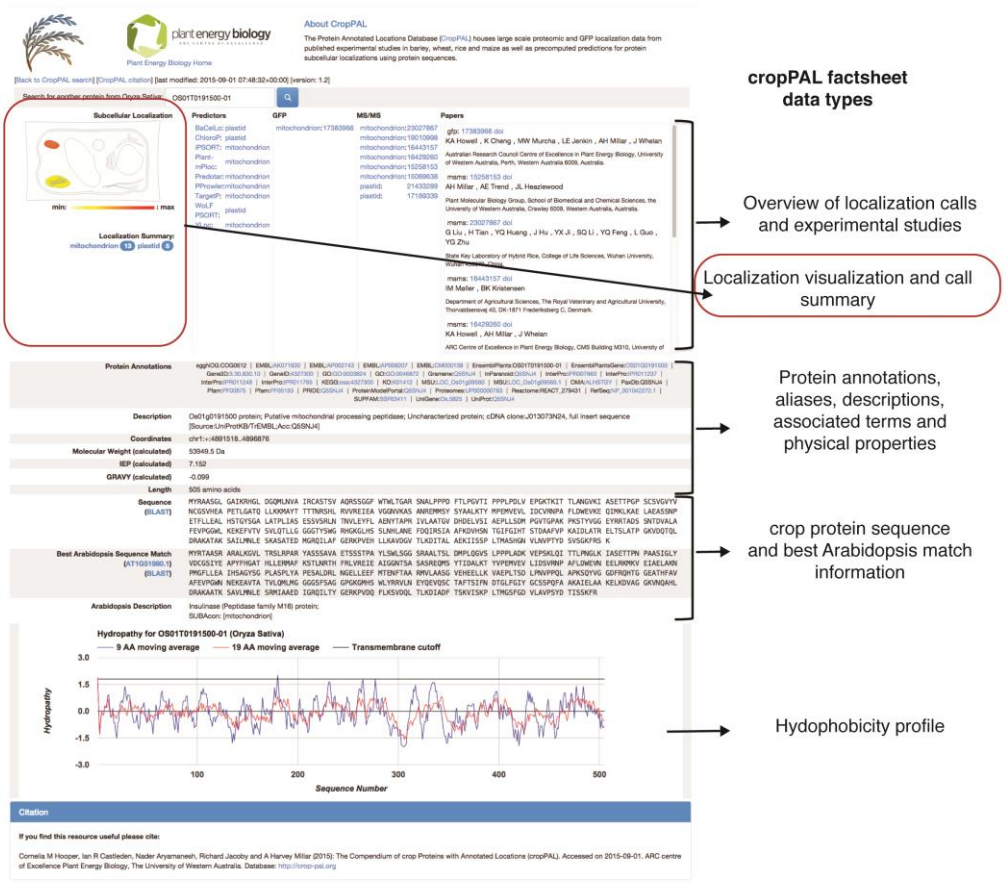
Fig. 2 The cropPAL interface. (A) The query interface allows choice of species and simple or complex building options as indicated by arrows. The Query builder tabs categorize the filter options into four categories (Subcellular Location, Protein Properties, Homology, Affiliations). (B) The results view offers a tabular overview of the protein hits fitting the query. The viewed information can be customized, downloaded or each protein hit can be investigated further by clicking on the protein identifier.

Fig. 3 The protein factsheet view. The collated information for each protein in cropPAL is available through the factsheet that can be opened *via* the hyperlinked identifiers in the results view. The localizations and a call count summary (highlighted in the red box), protein description are displayed as an overview. A number of information highlighted as blue hyperlinks lead to respective databases for more in-depth protein information.


Literature curation
GFP data
MS data
geographic
&
affiliation data
Identifier linking
Protein accessions
Uniprot, MSU, RAP, DFCI, NCBI
Sequence retrieval and
blast
Reference sequences
Gramene/Ensembl Plants
Reference genome, proteome
Annotations
Homology tree
User
Query builder
cropPAL interface
http://crop-pal.org
THE COMPENDIUM OF PROTEINS WITH ANNOTATED LOCATION
Predictor outputs
location predictions from 8 predictors
Interrogation
Cross-species gap filling
Sub-proteomes
Intersect proteomes
Protein families
etc.
SUBA3
(GFP, MS, predictors, classification)
TAIR10 - annotations


A
Crop species choice menu
ORYZA SAVITA
TRITICUM AESTIVUM
ZEA MAYS
HORDEUM VULGARE
Choose Rice
Choose Wheat
Menu Links
Tutorial
Localization World Map
cropPAL stats
Quick Search
(for text or protein identifier)
press buttons to obtain the results
Query category tab
Subcellular Location
Protein Properties
Filters in query tabs
add filter conditions to the query
The query window
(will take place of the Quick Search Window)
press query to obtain the results
Undo
Clear
Query


B
Tabular result view
Display columns can be customized
Query hit IDs
Result tab
each species and/or each query opens a separate result tab
query information and download option for results table
Best Arabidopsis match
Localization information
including literature references
Homology matches
Homologs in other crop species are displayed in the Homologies column as a hyperlink to the respective factsheet


cropPAL factsheet data types
Overview of localization calls and experimental studies
Localization visualization and call summary
Protein annotations, aliases, descriptions, associated terms and physical properties
crop protein sequence and best Arabidopsis match information
Hydophobicity profile