A Motif and Amino Acid Bias Bioinformatics Pipeline to Identify Hydroxyproline-Rich Glycoproteins[1][OPEN]

Kim L. Johnson[2], Andrew M. Cassin[2], Andrew Lonsdale, Antony Bacic, Monika S. Doblin[3], and Carolyn J. Schultz[3],*

Australian Research Council Centre of Excellence in Plant Cell Walls, School of BioSciences, University of Melbourne, Parkville, Victoria 3010, Australia (K.L.J., A.M.C., A.L., A.B., M.S.D.); and School of Agriculture, Food, and Wine, University of Adelaide, Waite Research Institute, Glen Osmond, South Australia 5064, Australia (C.J.S.)

ORCID IDs: 0000-0001-6917-7742 (K.L.J.); 0000-0002-0292-2880 (A.L.); 0000-0001-7483-8605 (A.B.); 0000-0002-8921-2725 (M.S.D.); 0000-0003-2026-9122 (C.J.S.).

Intrinsically disordered proteins (IDPs) are functional proteins that lack a well-defined three-dimensional structure. The study of IDPs is a rapidly growing area as the crucial biological functions of more of these proteins are uncovered. In plants, IDPs are implicated in plant stress responses, signaling, and regulatory processes. A superfamily of cell wall proteins, the hydroxyproline-rich glycoproteins (HRGPs), have characteristic features of IDPs. Their protein backbones are rich in the disordering amino acid proline, they contain repeated sequence motifs and extensive posttranslational modifications (glycosylation), and they have been implicated in many biological functions. HRGPs are evolutionarily ancient, having been isolated from the protein-rich walls of chlorophyte algae to the cellulose-rich walls of embryophytes. Examination of HRGPs in a range of plant species should provide valuable insights into how they have evolved. Commonly divided into the arabinogalactan proteins, extensins, and proline-rich proteins, in reality, a continuum of structures exists within this diverse and heterogenous superfamily. An inability to accurately classify HRGPs leads to inconsistent gene ontologies limiting the identification of HRGP classes in existing and emerging omics data sets. We present a novel and robust motif and amino acid bias (MAAB) bioinformatics pipeline to classify HRGPs into 23 descriptive subclasses. Validation of MAAB was achieved using available genomic resources and then applied to the 1000 Plants transcriptome project (www.onekp.com) data set. Significant improvement in the detection of HRGPs using multiple-*k*-mer transcriptome assembly methodology was observed. The MAAB pipeline is readily adaptable and can be modified to optimize the recovery of IDPs from other organisms.

Intrinsically disordered proteins (IDPs) challenge the traditional view that proteins fold into a fixed three-dimensional structure that determines their function (Babu, 2016). IDPs do not conform to the structure-function paradigm, as they are highly mobile, lack a persistent structure, and yet are stable (Schlessinger et al., 2011; Szalkowski and Anisimova, 2011; Forman-Kay and Mittag, 2013; Light et al., 2013). Fully sequenced eukaryotic proteomes suggest that IDPs are common, with 10% to 20% of proteins being completely disordered and 25% to 40% being partially disordered (Ward et al., 2004; Oates et al., 2013; Peng et al., 2014; Kurotani and Sakurai, 2015). The amino acid composition of IDPs is biased toward disorder-promoting residues, the majority of which are polar or charged, and Pro, which, despite being nonpolar, is the most disorder-promoting residue due to its rigid conformation. The resulting structure of IDPs lacks stable hydrophobic cores and is likely to expose most of their amino acids to solvent. IDPs also commonly contain sequence repeats and sequence motifs for recognition by enzymes that carry out posttranslational modifications (PTMs; Forman-Kay and Mittag, 2013). The accessibility of the protein backbones to these PTM enzymes and the effect of PTMs on the structural, steric, and electrostatic properties can result in IDPs having multiple binding partners. These properties make IDPs ideally suited for functions associated with transient molecular recognition; indeed, their important roles

[1] This work was supported by the Victorian Life Sciences Computation Initiative (grant no. VR0191) on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia; by the 1000 Plants Initiative, which is funded by the Alberta Ministry of Enterprise and Advanced Education, Alberta Innovates Technology Futures, Innovates Centre of Research Excellence, and Musea Ventures; and by the Australia Research Council Centre of Excellence in Plant Cell Walls (grant no. CE1101007 to K.L.J., A.M.C., A.L., M.S.D., and A.B.), which is funded by the Australian Government's National Collaborative Research Infrastructure Program administered through Bioplatforms Australia.

[2] These authors contributed equally to the article.

[3] These authors contributed equally to the article.

* Address correspondence to carolyn.schultz@adelaide.edu.au.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Carolyn J. Schultz (carolyn.schultz@adelaide.edu.au).

A.M.C. generated the HRGP Web site, and K.L.J., A.M.C., M.S.D., C.J.S., and A.B. designed and oversaw the research; A.L. provided bioinformatics support and assisted in maintenance of the HRGP Web site; K.J.L., A.M.C., M.S.D., and C.J.S. analyzed the data and, together with A.B., wrote the article.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.17.00294

in cellular signaling and regulation are becoming increasingly apparent.

Evolutionary studies suggest that IDPs have played an important role for progressing from simple to complex multicellular organisms (Dunker et al., 2015). As IDPs lack the sequence constraints required for maintaining a folded structure, they can display rapid evolution, providing a mechanism to increase regulatory complexity. However, IDPs often display high conservation of overall composition and specific motifs yet low sequence conservation due to high mutation rates, increased insertion/deletion events, and domain swapping (Buljan et al., 2010; Nido et al., 2012; Khan et al., 2015). These features make tracking IDPs over evolutionary time scales immensely challenging. Although a number of databases with either experimental data or predictive methods of protein disorder are available (for review, see Piovesan et al., 2017), the tools available to assess large-scale proteomics data for IDP evolution are lacking (Varadi et al., 2015). The methods available still rely on significant conservation of sequence order, as they utilize (PSI)-BLAST to retrieve sequences, a user-generated multiple sequence alignment, or knowledge of function (Varadi et al., 2015; Khan and Kihara, 2016). Since IDPs have strong amino acid biases but relatively low sequence similarity, approaches such as standard BLAST searches are not effective for identification over long evolutionary distances. In order to capitalize on current and emerging genomics and transcriptomic resources, the development of novel bioinformatics approaches to identify IDPs from any organism would represent a significant advance.

In plants, IDPs are implicated in stress responses, signaling, and molecular recognition pathways (Pietrosemoli et al., 2013; Sun et al., 2013). As plants are sessile, IDPs related to environmental stress responses are proposed to be particularly critical to enable adaptation to challenging environments (Gomord et al., 2010; Pietrosemoli et al., 2013; Kurotani and Sakurai, 2015). An important class of extracellular IDPs that are believed to be involved in these responses are the hydroxyproline-rich glycoproteins (HRGPs), an evolutionarily ancient and diverse family of cell wall proteins. The protein backbones of HRGPs consist of different Pro-rich motifs that govern Pro hydroxylation and direct subsequent HRGP glycosylation. It is these features of HRGPs that define them as IDPs.

HRGPs have been detected and isolated in both the protein-rich walls of green algae and the cellulose-rich walls of land plants (Supplemental Table S1). Proteins, including glycoproteins, can be significant components of some algal walls (e.g. *Chlamydomonas reinhardtii*, 30% [w/w]; Roberts, 1974; Voigt et al., 2009), yet they are generally a minor component of land plant (embryophyte) walls (approximately 10% [w/w] in primary walls but less in secondary walls). Despite their low abundance in land plants, selected HRGPs have been shown to play important functional roles in, but not limited to, cell expansion, root growth and development, xylem differentiation, somatic embryogenesis, initiation of female gametogenesis, self-incompatibility, signaling, salt tolerance, and pathogen responses, (for review, see Fincher et al., 1983; Kieliszewski and Lamport, 1994; Majewska-Sawka and Nothnagel, 2000; Seifert and Roberts, 2007; Ellis et al., 2010; Lamport et al., 2011; Draeger et al., 2015; Velasquez et al., 2015; Showalter and Basu, 2016a, 2016b). Not surprisingly, HRGPs have both fascinated and challenged researchers for decades.

The wide range of potential functions of these molecules lies in the diversity of protein backbones and the vast potential for PTMs. HRGPs are commonly divided into three major multigene families, the arabinogalactan proteins (AGPs), extensins (EXTs), and Pro-rich proteins (PRPs), but in reality, a continuum of HRGPs exists, including hybrids with features from multiple HRGP families, chimeric HRGPs that contain additional non-HRGP protein domains, and very small proteins, such as the arabinogalactan (AG)-peptides (Fig. 1; Johnson et al., 2003b). The common thread that defines this diverse family is the hydroxylation of Pro to Hyp (O) and the subsequent attachment of *O*-linked glycans on Hyp residues. Glycosylation of Hyp is a widespread phenomenon in plants but is absent in animals. Plant HRGPs have been considered functionally equivalent to mammalian proteoglycans such as mucins that are rich in Pro, Thr, and Ser, heavily glycosylated, and found in the extracellular matrix (Chaturvedi et al., 2008). Importantly, however, *O*-linked glycosylation of mucins occurs through Ser and Thr residues rather than Hyp residues, which are rarely, if ever, glycosylated in animals.

The HRGPs range from highly glycosylated molecules, such as the AGPs, to the moderately glycosylated EXTs and minimally glycosylated PRPs. Analysis of the Hyp-containing motifs that direct this glycosylation led to the Hyp-contiguity hypothesis. This predicts that small glycan structures such as arabinooligosaccharides (arabinosides; degree of polymerization, 1–4) are attached to contiguous Hyp residues such as the $SO_{3\text{-}5}$ repeats that occur in EXTs (Fig. 1). In the highly glycosylated AGPs, large type II arabino-3,6-galactan polysaccharides (degree of polymerization, 30–120) are added to clustered, non-contiguous Hyp residues such as the SO, AO, TO, and VO repeats (Fig. 1; Kieliszewski and Lamport, 1994; Lamport et al., 2011). The extent of glycosylation of PRPs remains unclear, as it has not been studied extensively. For example, purified soybean (*Glycine max*) PRPs were shown to be minimally glycosylated, constituting less than 3% carbohydrate, with Ara residues presumably linked to Hyp (Datta et al., 1989; Fig. 1).

HRGPs represent a challenging group of IDPs, given that they consist of large multigene families with diverse protein backbones (Schultz et al., 2002; Showalter et al., 2010). Examination of these glycoproteins in a wide range of plant species should provide valuable insights into how such proteins have evolved. In this article, we use the HRGPs as a test case to develop a versatile tool to identify and classify specific subsets of

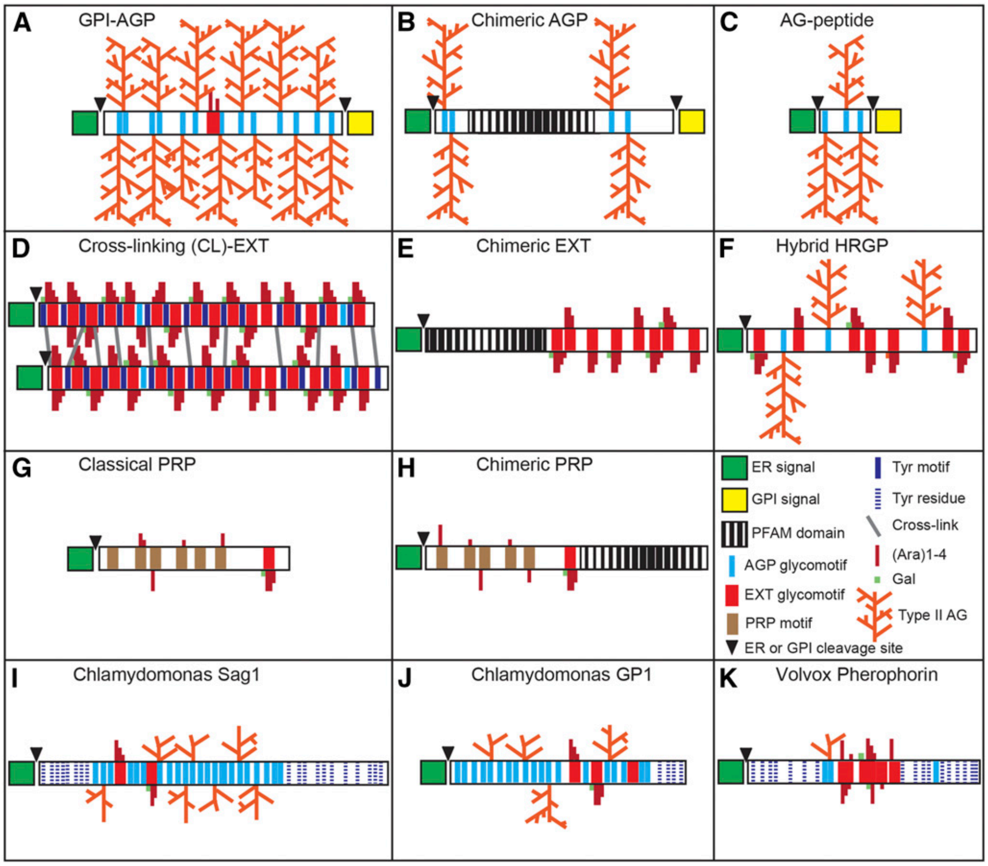


Figure 1. Schematic of the predicted structures of selected HRGPs. The major angiosperm HRGP multigene families are the AGPs (A–C), cross-linking (CL)-EXTs (D and E), and PRPs (G and H). Hybrid HRGPs (F) contain motifs characteristic of more than one HRGP family and are commonly found in green algae (I–K). The protein motifs that direct the hydroxylation of Pro to Hyp and undergo subsequent *O*-glycosylation are as follows: (1) SP, AP, TP, VP, and GP (light blue bars), to which large type II arabinogalactan chains (type II AG; orange) are added; (2) SP_{3-5} glycomotif repeats (red bars) that direct the addition of short arabinose (Ara) side chains (dark red) on Hyp residues and Gal (green) on Ser residues. In CL-EXTs, these SP_{3-5} motifs alternate with Y cross-linking motifs (dark blue bars representing YXY, VYK, and YY) in the protein backbone. Y motifs can form both intramolecular and intermolecular cross-links. Intermolecular cross-links (gray) occur through the formation of diisodityrosine. Algal HRGPs have single Tyr residues outside the Pro-rich regions (dark blue, dashed; I–K); (3) PRP motifs (brown bars) direct minimal glycosylation of short Ara residues (G and H). Chimeric HRGPs have a recognized PFAM domain (black vertical lined box; B, E, and H) in addition to a HRGP region.

defined IDPs. Relatively little is known about the origin and evolution of HRGPs, as evolutionary studies have largely focused on the proteins involved in the synthesis and remodeling of the major polysaccharides, cellulose, the noncellulosic polysaccharides (hemicelluloses), and pectins. As most of the HRGPs from chlorophyte algae have domain structures distinct from embryophyte HRGPs (Fig. 1; Godl et al., 1997; Ferris et al., 2005), a tool that is able to identify HRGPs with diverse features across long evolutionary time scales was required.

Bioinformatics approaches to identifying HRGP family members have been developed and proven useful to characterize this complex family. The protein backbone of AGPs is rich in the amino acids Pro/Hyp, Ala, Ser, and Thr of PAST (Schultz et al., 2002; Ma and Zhao, 2010; Showalter et al., 2010). A biased amino acid approach (50% or greater PAST) was developed to detect Arabidopsis (*Arabidopsis thaliana*) AGPs (Schultz et al., 2002), which also identified some EXTs and PRPs (Johnson et al., 2003b). Showalter et al. (2010) modified this approach, developing BIO OHIO to include a mix of different strategies: a biased amino acid approach was used for AGPs (50% or greater PAST), a motif-based search for EXTs (two or more SP_4 [or SP_3]) motifs, and a combination of both methods for PRPs. Many EXTs also include Y-based CL motifs, and searching for SP_{3-5} motifs identifies both EXTs with Y motifs (CL-EXTs) or without (non-CL-EXTs). PRPs are the most difficult class of HRGPs to define (Johnson et al., 2003b; Showalter et al., 2010). Arabidopsis has two distinct sub-classes of PRPs, one rich in PTYK (e.g. *AtPRP1* and *AtPRP3*) and a second rich in PVKC (e.g. *AtPRP2* and *AtPRP4*; Fowler et al., 1999). Members of both subclasses are chimeric

(Johnson et al., 2003b), containing a PFAM domain like those identified in *Populus trichocarpa* using the revised BIO OHIO 2.0 program (Showalter et al., 2016). Thus, Showalter et al. (2010, 2016) used a combination of both bias and motifs to search for PRPs [45% or greater PVKCYT or 2 or more KKPCPP motifs or 2 or more PPVX(K/T) motifs] in Arabidopsis and *P. trichocarpa*. With these approaches (Showalter et al., 2010, 2016; Liu et al., 2016), many HRGP sequences were placed into more than one HRGP family. Therefore, manual curation of BIO OHIO output is required for final classification (Showalter et al., 2010, 2016), a task that is not practicable with large data sets and that also has the potential to introduce subjective bias.

Another, less biased approach based on a clustering method was used by Newman and Cooper (2011) to search for tandem repeats to identify Pro-rich proteins in plants. While EXTs were able to be identified, this method was poor in detecting AGPs because their diagnostic AP, SP, and TP glycomotifs that direct arabino-3,6-galactan-type glycosylation (Tan et al., 2003) are scattered throughout the protein backbone, rather than being found in tandem (Schultz et al., 2002; Showalter et al., 2010). The method identified three AGP-associated motifs, TPPPA, MTPPS, and MTPPPMP, which are found in only a few AGPs.

This study builds on these previous bioinformatics approaches to identify and classify HRGPs in 15 plant genomic data sets available through Phytozome and within the extensive 1000 Plant (1KP) transcriptome data set (Johnson et al., 2017). The 1KP project (www.onekp.com) was established to allow researchers to more effectively address evolutionary and other questions by providing deeper sampling of key plant taxa beyond those with sequenced genomes (Johnson et al., 2012; Matasci et al., 2014; Wickett et al., 2014; Xie et al., 2014).

Here, we present a novel, highly robust, and automated motif and amino acid bias (MAAB) pipeline to classify sequences into one of 24 predefined descriptive sub-classes (23 HRGP classes and one MAAB class). This approach combines a biased amino acid approach with a relative motif-counting step. We optimized MAAB to identify GPI-AGPs, non-GPI-AGPs, and CL-EXTs. PRPs and hybrid-type HRGPs that reflect variants of each of the major classes also were captured. The pipeline is publicly available (http://services.plantcell.unimelb.edu.au/hrgp/index.html) and readily adaptable for the identification of other IDPs from plants and other organisms. Because chimeric HRGPs, such as fasciclin-like AGPs (Johnson et al., 2003a), lipid-transfer proteins (Motose et al., 2004), and AG-peptides (Schultz et al., 2004), also can be identified by other approaches, they will be reported elsewhere. Here, we provide a detailed analysis of HRGPs from 15 plant genomes, including two volvocine algae, a moss, a lycophyte, four commelinid monocots, a basal eudicot, and six core eudicots. We demonstrate the importance of using a multiple-*k* assembly approach to recover HRGPs from the 1KP transcriptome data and provide preliminary analysis of 1KP data to validate this approach. The biological and evolutionary analysis of the 1KP data are provided in the companion article (Johnson et al., 2017).

RESULTS AND DISCUSSION

Arabidopsis GPI-AGPs and CL-EXTs Are Distinct Types of IDPs

There are two major challenges in undertaking a bioinformatics search of genomic and transcriptomic data sets for HRGPs: (1) initial robust identification, and (2) establishment of unambiguous criteria for subsequent classification. These issues are substantive and challenging given the enormous diversity of HRGP members even within any one subfamily of any particular plant species. For example, of the 85 AGPs identified with BIO OHIO/manual curation (Showalter et al., 2010) in Arabidopsis, six different sub-classes were defined including classical, Lys-rich classical, AG-peptide, and three subclasses of chimeric AGPs (fasciclin-like AGPs, plastocyanin AGPs, and others). The intrinsically disordered mature protein core of AGPs and EXTs can clearly be seen when HRGP sequences are assessed using three different disorder predictors (PONDR-VL-XT, VL3, and VSL2; Romero et al., 1997). AtAGP6 and AtEXT3 show high disorder scores (greater than 0.5) along the entire length of the mature protein sequence compared with the order seen in a typical globular protein such as prolyl-4-hydroxylase (Fig. 2A). Some features can be observed in IDPs; for example, the default VL-XT predictor shows the repetitive nature of AtEXT3 (Fig. 2A).

The diversity of GPI-AGPs, even within a single plant species, is highlighted by an alignment of the 16 Arabidopsis GPI-AGPs (Fig. 2B). These GPI-AGPs were identified primarily by amino acid bias (Schultz et al., 2002; Showalter et al., 2010), and shading of motifs is used to highlight the differences and similarities between members (Fig. 2B). Hereafter, we refer to HRGP glycosylation motifs as glycomotifs and use SP, not SO (except where protein sequencing has been performed), since we are working with proteins predicted from genomes/transcriptomes. Thus, the hydroxylation (and, therefore, glycosylation) status of Pro is unknown and context dependent (Tan et al., 2003; Shimizu et al., 2005; Kurotani and Sakurai, 2015). The relatively low sequence similarity of members is clearly evident when viewing a percentage identity matrix, where GPI-AGPs range from 9.6% to 61.6% and CL-EXTs range from 6.3% to 84.2% amino acid identity (Supplemental Fig. S1, A and B, respectively). The diversity of GPI-AGPs and CL-EXTs is further highlighted by phylogenetic analyses (Fig. 2, C and D). For the 16 Arabidopsis GPI-AGPs, there are some robust groupings (greater than 70% bootstrap values), but only between pairs of sequences. For example, the pollen-specific AtAGP6 and AtAGP11 (Pereira et al., 2006; Coimbra et al., 2009, 2010) and the Lys-rich AtAGP17 and AtAGP18 (Gaspar et al., 2004;

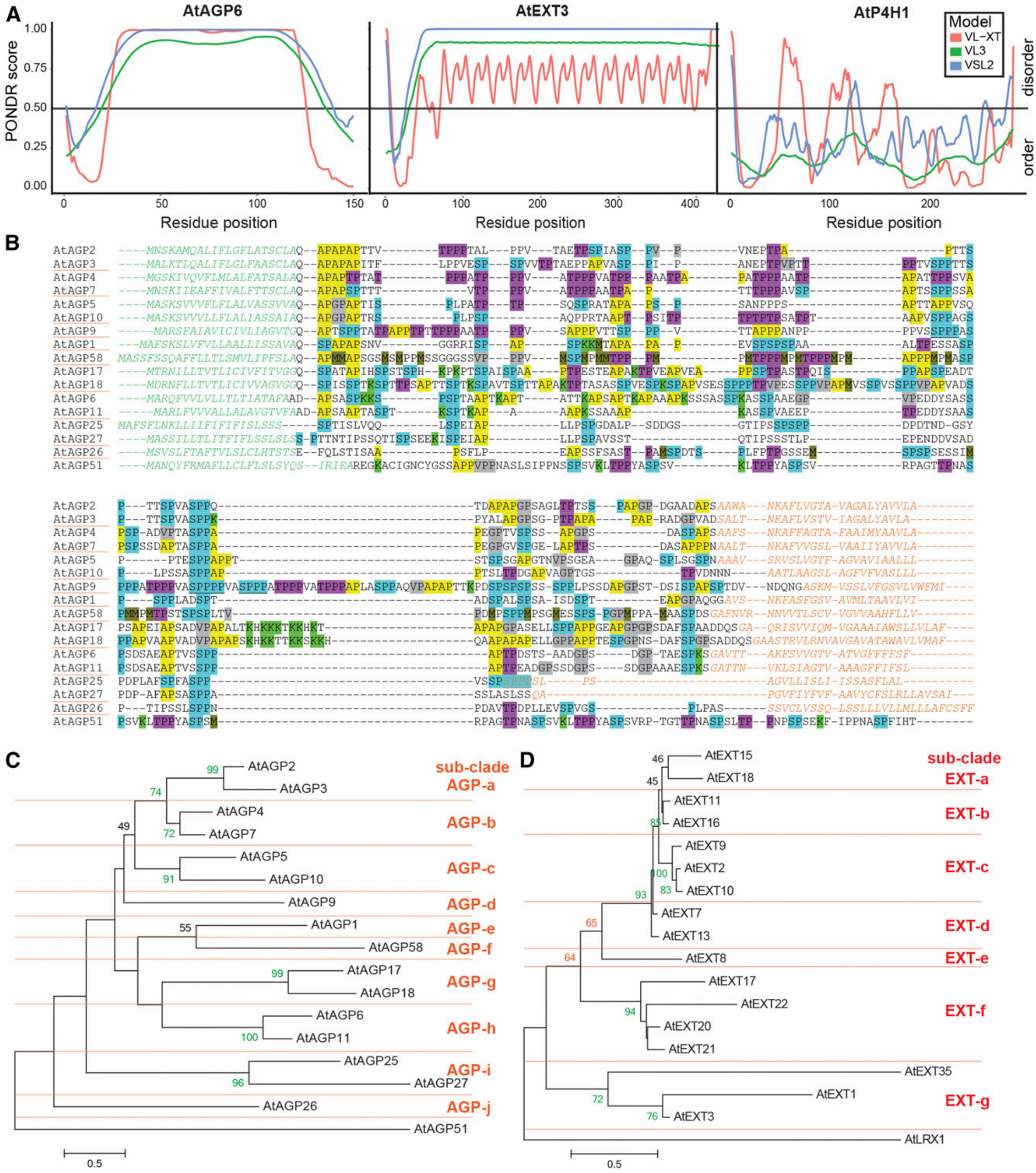


Figure 2. Disorder prediction, sequence alignment, and phylogenetic trees of the Arabidopsis classical GPI-AGPs and CL-EXTs. A, Protein disorder (PONDR) plots (see "Materials and Methods") for AtAGP6 (*At5g14380*; left), AtEXT3 (*At1g21310*; middle), and prolyl-4-hydroxylase (AtP4H1; *At2g43080*; right) using VL-XT (red), VL3 (green), and VSL2 (blue). PONDR prediction scores above the threshold line (0.5) predict disorder; below the line, they predict order. B, Sequence alignment (MUSCLE) of 16 Arabidopsis GPI-AGPs with the non-GPI-AGP AtAGP51 included for comparison. Endoplasmic reticulum (ER; N-terminal) and GPI-anchor (C-terminal) signal sequences are colored in green and orange, respectively. Glycomotifs and selected residues are highlighted as follows: $AP_{1\text{-}3}$ (yellow); $SP_{1\text{-}2}$ (blue); SPPP (also found in EXTs; blue underlined); $TP_{1\text{-}3}$ (pink/purple); $[G/V]P_{1\text{-}3}$ (gray); K (bright green); and M (olive green). This shows the diversity and lack of sequence conservation between family members. C, Maximum likelihood tree (MEGA) of Arabidopsis GPI-AGPs and AtAGP51. D, Maximum likelihood tree (MEGA) of Arabidopsis CL-EXT and AtLRX1 (chimeric CL-EXT). In C and D, numbers on the nodes represent support with 100 bootstrap replicates

Yang et al., 2007, 2011) share 51% and 44.3% amino acid identity, respectively (Supplemental Fig. S1A). In total, 10 GPI-AGP subclades (AGP-a to AGP-j) can be identified; however, low bootstrap support is generally observed between them (Fig. 2C).

EXTs in Arabidopsis are similarly diverse, with Showalter et al. (2010) describing nine subclasses depending on the type of repeat (SP_3, SP_4, SP_5, or a combination), short, and chimeric (LRXs, PERKs, and others). Of the 59 EXTs reported by Showalter et al. (2010), we redefined SP_3, SP_4, and SP_5 EXTs and one short EXT (EXT35) as CL-EXTs, as they satisfied the criteria of containing a signal peptide, at least two YXY motifs, and no GPI-anchor signal (see "Materials and Methods"). Based on these requirements, we compared 16 CL-EXT sequences using a percentage identity matrix. Although pairwise similarity is generally higher than for the GPI-AGPs (compare Supplemental Fig. S1, A and B), the multiple sequence alignments highlight the range of sequence lengths, diversity, and spacing of $SP_{3\text{-}5}$ motifs and Y-based cross-linking motifs that separate them and the presence of other glycomotifs such as the AGP-like SPSP motifs (Supplemental Fig. S2; Saha et al., 2013). Nine of the Arabidopsis CL-EXTs form a well-supported clade (93% bootstrap support) comprising four subclades (EXT-a to EXT-d), with an additional eight CL-EXTs placed in three other subclades (EXT-e to EXT-g; Fig. 2D). AtEXT17/22/20/21 form a robust subclade (EXT-f); however, poor bootstrap support is observed between other CL-EXT family members (Fig. 2D), similar to our analysis of GPI-AGPs.

The low sequence similarity and diversity of classical GPI-AGPs and CL-EXTs makes them extremely difficult to detect using BLASTp, as, even within the rosids, it was not possible to identify all of the putative Arabidopsis orthologs (Table I; Showalter et al., 2010); hence, the need for a new, more robust search tool.

MAAB Pipeline Construction and Validation for the Identification and Classification of HRGPs Using 15 Predicted Proteomes

The MAAB pipeline (summarized in Fig. 3; see "Materials and Methods") was created to identify and classify HRGPs in any given proteome without an excessively high level of false positives. As previous bioinformatics studies of HRGPs have largely been undertaken in Arabidopsis (Schultz et al., 2002; Showalter et al., 2010), the MAAB pipeline was initially parameterized in an iterative manner on this proteome to optimize the recovery and appropriate classification of the HRGPs identified in this study. Additional features were then incorporated into MAAB based on studies of HRGPs in algal species (Godl et al., 1997; Ferris et al., 2005). MAAB can robustly identify HRGPs, as it incorporates both amino acid biases and protein motif analysis, in two stages: (1) finding all HRGPs and removing AG-peptides and all chimeric HRGPs; and (2) classification, including primary classification based on amino acid bias, motif analysis, and final classification (Fig. 3; see "Materials and Methods"). Key steps in stage 1 of the pipeline are removal of likely duplicates (1a), a requirement for 45% or greater PAST, PSKY, or PVKY (1b), a length threshold to reduce the number of partial sequences (1c), a 10% or greater Pro filter (1d), removal of chimeric HRGPs with PFAM domains (1e), and a requirement for an ER signal sequence (1f). Subsequent steps in stage 2 in the MAAB pipeline, including prediction of GPI-anchor addition to the C terminus, allowed us to identify and distinguish between the major categories of interest, GPI-AGPs, CL-EXTs, and PRPs, and capture potentially different HRGPs in algal and non-vascular plant lineages. All non-chimeric (classical and hybrid) HRGP sequences were categorized into one of 23 unique HRGP classes, and a final class, MAAB class 24, contains predominantly non-HRGPs (or unknown HRGPs) based on having less than 15% known HRGP motifs (Fig. 3; see "Materials and Methods").

The output of the MAAB pipeline, using the predicted proteome data sets from 15 completed land plant and algal genomes (see "Materials and Methods") available at Phytozome (Goodstein et al., 2012), is summarized in Table I. The full data matrix, including sequences, percentage amino acid bias, and motif counts, is provided in Data File 1. HRGPs were identified in all proteomes with the exception of the picoplankton *Ostreococcus lucimarinus*. In most eudicots, the majority of HRGP sequences (classes 1–23), accounting for between 52% (*Solanum lycopersicum*) and 87% (*Eucalyptus grandis*) of the total hits, fall into three HRGP classes: GPI-AGPs (class 1), CL-EXTs (class 2), and non-GPI-AGPs (class 4). The other major class is MAAB class 24 (less than 15% known HRGP motifs; Table I). The MAAB pipeline identified as many or more GPI-AGPs and CL-EXTs than the number detectable by BLASTp, particularly outside eudicots (Table I). This indicates that MAAB is successful at capturing and classifying GPI- and non-GPI-AGPs and CL-EXTs (classes 1, 4, and 2, respectively), hybrid and potentially unknown HRGPs (classes 5–23), as well as other biased amino acid proteins, most of which are not HRGPs (MAAB class 24; discussed in detail below).

Motif shading of the sequences identified by MAAB further supports the correct classification of HRGPs in classes 1 to 23 and the low number of known HRGP

Figure 2. (*Continued.*)
(70 or greater, green; 60–69, orange, 40–59, black) with subclades AGP-a to AGP-j (C) and EXT-a to EXT-g (D; denoted by horizontal lines). Scale bars for branch length measure the number of substitutions per site. The CL-EXT alignment with shaded motifs is shown in Supplemental Figure S2.

Table I. *Comparison of the number of HRGPs identified by BLASTp (National Center for Biotechnology Information [NCBI]) and MAAB in Phytozome genomes*

Species/Clade	BLASTp[a]		MAAB Class[b]																				Totals		
	GPI-AGP	CL-EXT	1	2	3	4	5	6	7	9	11	12	15	16	17	18	19	20	21	22	23	24	Total Class 1–4	Total Class 5–23	Total Class 1–24
			GPI-AGP	CL-EXT	PRP	Non-GPI-AGP	AGP Bias			GPI-EXT	EXT Bias		PRP Bias				Shared Bias					<15% Motif			
Monocots/commelinids																									
Brachypodium distachyon	1		14			20									3							11	34	3	48
Oryza sativa	4		11			18	2		1			1										38	29	4	71
Setaria italica	2		6			13	1								1							13	19	2	34
Sorghum bicolor	3		9			24	2								3		3					18	33	8	59
Core eudicots/asterids																									
Solanum lycopersicum	8	7	11	9		15	3	2				19					4	2		2		8	35	32	75
Core eudicots/rosids																									
Arabidopsis thaliana	16	16	17	16		12	1			1	1					1		7	2		1	7	45	14	66
Eucalyptus grandis	3		2			5												1				18	7	1	26
Glycine max	7		20	5	1	17	3	3								7	2	1				24	43	16	83
Medicago truncatula	8	3	13	3		14	1						1				4					8	30	6	44
Populus trichocarpa	12		10	3		13	1				1						1	4	2	1		23	26	10	59
Basal eudicots																									
Aquilegia coerulea			4			6											1					6	10	1	17
Lycophytes																									
Selaginella moellendorffii	1	5	3	9		9	1	1				1		1			1	1				2	21	6	29
Mosses																									
Physcomitrella patens	5		12			20											3					1	32	3	36
Green algae																									
Volvox carteri			1			5	2										1	1				8	6	4	18
Chlamydomonas reinhardtii			2			78	1	2									2					28	80	5	113

[a] Arabidopsis GPI-AGPs (16) and CL-EXTs (16) from Showalter et al. (2010; see Supplemental Table S2) were used as query sequences (NCBI; word size = 2, no filtering) to identify AGPs/EXTs in each organism. [b] The number of sequences identified by MAAB is shown by HRGP class for each species, and no sequences were detected for HRGP classes 8, 10, 13, and 14.

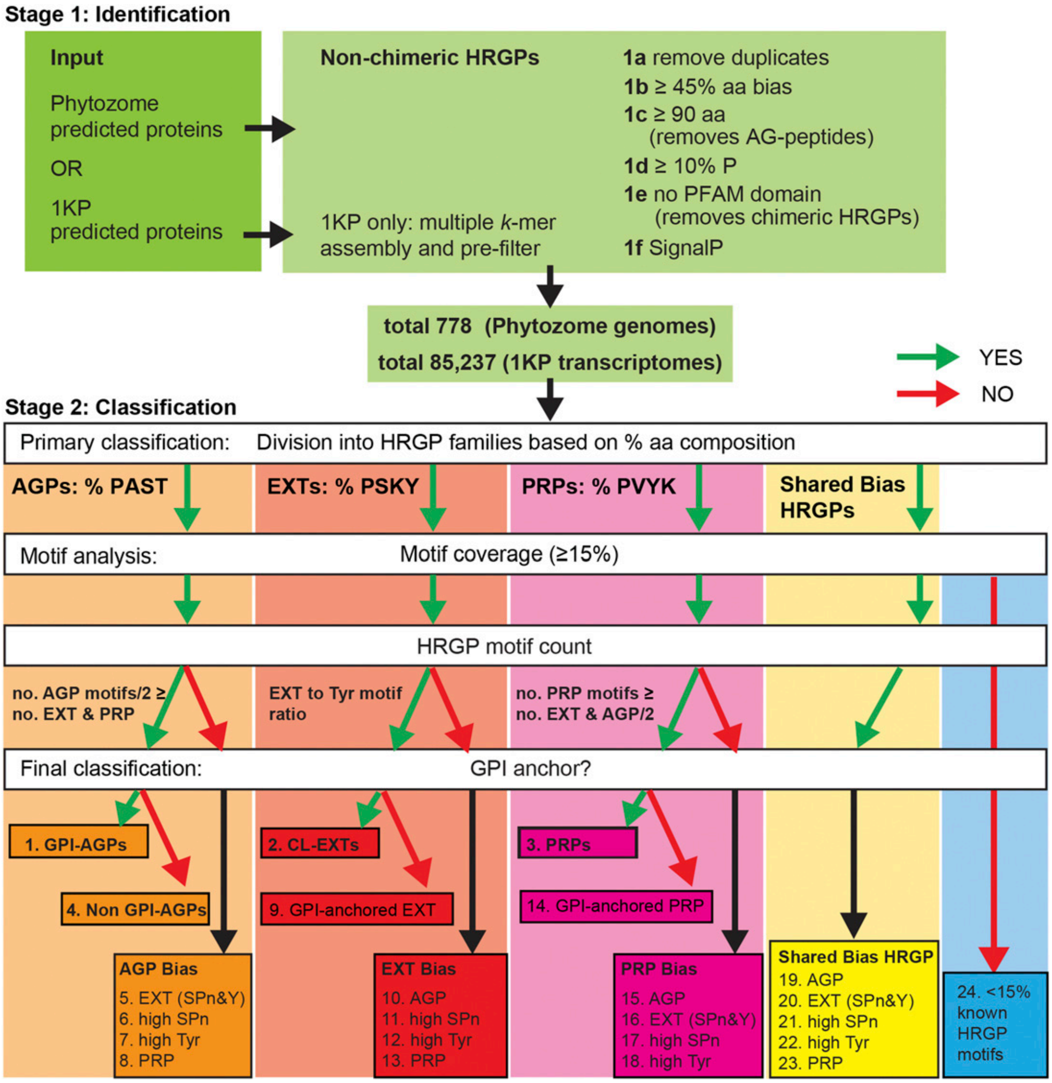
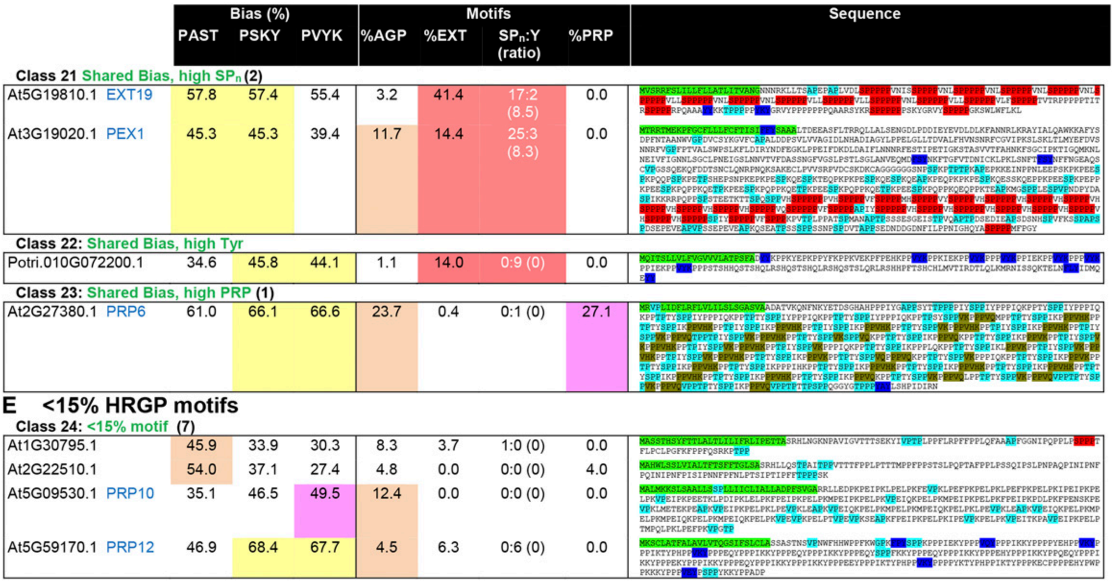


Figure 3. Overview of the MAAB pipeline for the identification and classification of non-chimeric HRGPs. The pipeline consists of two major stages: stage 1 (1a–1f), identification; and stage 2, classification. Stage 1 largely consists of removing unwanted sequences, including chimeric HRGPs and AG peptides, and retaining sequences with the desired amino acid bias (45% or greater) and ER signal sequence. Stage 2 filters sequences into four categories based on the percentage amino acid composition that is dominant by 2% or greater: AGPs (boxed in orange) if PAST, EXTs (boxed in red) if PSKY, and PRPs (boxed in pink) if PVKY. If no clear bias exists (Δ amino acid bias < 2%) the sequence is placed in the shared bias HRGPs (boxed in yellow). The next step is HRGP motif analysis, which uses motif type and number (no.). The motifs used for AGPs are [ASVTG]P, [ASVTG]PP, [AVTG]PPP; those used for EXT are SP_3, SP_4, SP_5, [FY]XY, KHY, VY[HKDE], VxY, and YY; and those used for PRPs are PPV[QK], PPVx[KT], and KKPCPP. A relative HRGP motif count (for AGP and PRP bias) ensures that sequences have the motifs expected for the amino acid bias class they are categorized into (see "Materials and Methods"). The number of accepted AGP motifs is calculated from the number of AGP motifs divided by 2 (since two typical AGP motifs [e.g. SPAP] have a similar length to a typical EXT motif [e.g. SPPP] and a typical PRP motif [e.g. PPVxK]). Accepted CL-EXT motifs have a minimum requirement of two SP_{3-5} motifs and two Y motifs that must be present in a similar ratio (SP_n:Y between 0.25 and 4). An additional MAAB class (class 24) arises for proteins with less than 15% known HRGP motifs (boxed in blue). After HRGP motif classification, the sequences that do not meet the above criteria (red arrow) are analyzed separately from the classical classes and placed into classes representing hybrid HRGPs. Before the final classification, all sequences are analyzed for the presence of a C-terminal GPI-anchor signal sequence. Sequences are thus categorized into one of 24 classes (Table I; see Fig. 4) with 23 classes of HRGPs: classes 1 to 4 representing the classical HRGPs classes; classes 5 to 23 representing minor HRGP classes consisting of, for example, hybrid HRGPs; and a final class, MAAB class 24, likely representing either non-HRGPs or unknown HRGPs.

motifs in MAAB class 24 (Supplemental Fig. S3). Selected sequences and parameters used for MAAB classification are illustrated in Figure 4. The classical GPI-AGPs (class 1) and non-GPI-AGPs (class 4) have AG-glycomotifs scattered through the entire mature protein backbone (after cleavage of ER and GPI signal sequences) and are distinct from the other sequences in the AGP bias classes 5 to 8 (Fig. 4A) that have different

	Bias (%)			Motifs				Sequence
	PAST	PSKY	PVYK	%AGP	%EXT	SP$_n$:Y (ratio)	%PRP	
A AGP Bias								
Class 1: GPI-AGPs (17)								
At5G14380.1 AGP6	64.7	39.3	28.7	30.0	0.0	0:0 (0)	0.0	MARQFVVLVLLTLTIATAFAADAPSASPKKSPSPTAAPTKAPTATTKAPSAPTKAPAAAPKSSSASSPKASSPAAEGPVPEDDYSASSPSDSAEAPTVSSPPAPTPDSTSAADGPSDGPTAESPKSGAVTTAKFSVVGTVATVGFFFFS
At3G27416.1 AGP59	67.3	43.9	24.6	29.2	0.0	0:0 (0)	0.0	MTLKLLSLAVAALAITAVLAADPPTPANAPKANGEKNSTSPSPAAAASPKSPTASAPTSPPTAAPTMAKKNSTGTPSPSPSSPSPKSSSAKTPASSPDSSSGDSSEGPTSSSDAPTASSPPAPTPEMSPSSDDGTGASDGPEASAPAAGASSLVISVSGSVLAAVAWLFFI
Class 4: Non-GPI-AGPs (12)								
At1G31250.1 AGP51	54.5	45.5	33.3	28.5	0.0	0:0 (0)	0.0	MANQYFRMAFLLCLFLSLSYQSIRIEAREGKACIGNCYGSSAPPVPPNASLSIPPNSSPSVKLTPPYASPSVKLTPPYASPSVRPAGTTPNASPSVKLTPPYASPSMRPAGTPNASPSVKLTPPYASPSVRPTGTTPNASPSLTPPNPSPSEKFIPPNASPFIHT
At1G68725.1 AGP19	68.1	45.2	39.9	31.9	11.3	7:0 (0)	2.0	MESNSIIWSLLLASALISSFSVNAQGPAASPVTSTTTAPPPTTAAPPTTAAPPPTTTTPPVSAAQPPASPVTPPPAVTPTSPPAPKVAPVISPATPPPQPPQSPPASAPTVSPPPVSPPPAPTSPPPTPASPPPAPASPPPAPASPPPAPVSPPPVQAPSPISLPPAPAPAPTKHKRKHKHKRHHHAPAPAPIPPSPPSPPVLTDPQDTAPAPSPNTNGGNALNQLKGRAVMWLNTGLVILFLLAMTA
Class 5: AGP Bias, high EXT (SP$_n$ & Y) (1)								
At1G02405.1 EXT30	47.8	44.0	41.0	6.0	14.9	3:2 (1.5)	0.0	MSPKTRRSTDLAATILMAIVILASPIIINAEDSSAEVDVNCIPCLQNQPPPPPSPPPPSCTPSPPPPSPPPPKKSSCPPSPLPPPPPPPPPNYVYTYPPGDLYPIENYYGAAVAVESFSVMKLFVFGVMVFLIL
Class 6: AGP Bias, SP$_n$								
Glyma01g45440.1	72.0	54.5	41.2	24.6	28.4	15:0(0)	0.0	METHLVLLLALIFTVVAGVGAQGPSTSPAPPTPQSSPPPIQSSPPPVPSSPPPAQSPPPASTPPPAPLSSPPPASPPPSSPPPASPPPASPPPASPPPASPPPASPPPASPPPATPPPASPPPFSPPPATPPPATPPPALTPTPLSSPPATSPAPAPAKVVAPALSPSLAPGPSLSTISPSGDDSGAEKLWSFQKMIGSLVFGCALLSLLF
Glyma02g01850.1	54.1	50.2	46.4	12.9	14.8	6:1(6)	0.0	MAPFPLMCFVILLLSSTVSVTATATQTTLTVGKLDQVPHIMCKECELPPPPPKECPPPPIPLAPPPPLPPSPPPPLPPSPPPPLPPSPPPPLPPPIPLSPPPPLPPSPPPPLPPSPSPPNLPSPPPLAPPPPLPPLAPPPPPLPPLPIPPNGPGEYLVPPPRDPNLPDLDDLNGAKIINPLLCFSNTSYFSLHSFTLLVLLCLPYYFFM
Class 7: AGP Bias, high Tyr								
LOC_Os01g02160.1	51.6	42.8	36.5	11.3	6.3	0:4 (0)	0.0	MASYRLLVILVFSALVALAAGDTYPADCPYPCLLPPPTPVTTDCPPPPSTPSSGYSYPPPSSSSSNTPPSSSSYWNYPPPQGGGGGYIPYYQPPAGGGGGGGYNYPAPPPPNPIVPWYPWYYRSPPSSPATAVTARGRSLLASVAVVTAAAAALITVF
Class 8: AGP Bias, high PRP								
UZWG_Locus_470	59.1	50.9	51.8	15.5	13.6	3:1 (3)	13.6	MALNAMLFLFTTFLVFTATVSSKDEELKEVTYVKEPQPAQPPAKVPVYAPPPPVKTPPSQSPPPKAPAYAPPPPPVNTPPSQSPPPKAPAYAPPPPPVNTPPSQSPPPKA
B EXT Bias								
Class 2: CL-EXTs (16)								
At2G43150.1 EXT8	60.8	76.4	70.8	0.9	68.9	22:12 (1.8)	0.0	MATPAWSHAKAQWVVAMLALLVGSAMATEPYYYSSPPPPYEYKSPPPPVKSPPPPYEYKSPPPPVKSPPPPYYYHSPPPPVKSPPPPYVYSSPPPPVKSPPPPYYYHSPPPPVKSPPPPYYYHSPPPPVKSPPPPYYYHSPPPPVKSPPPPYYYHSPPPPVKSPPPPYYYHSPPPPVKSPPPPYLYSSPPPPVKSPPPPVYIYASPPPPTHY
At1G26240.1 EXT20	63.2	85.6	78.5	2.1	80.5	43:44 (1)	0.0	MANPNGWPSLLMLVIALYSVSAHTSAQYTYSPPSPPSYVYKPPTHIYSSPPPPPYVYSSPPPPPYIYKSPPPPPYVYSSPPPPPYIYKSPPPPPYVYSSPPPPPYIYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYNSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYNSPPPPPYVYKSPPPPPPYVYSSPPPPSPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPSPPPPYVYKSPPPPPSYSYSYSSPPPPIY
Class 9: GPI-anchored EXT (1)								
At3G06750.1 EXT34	40.1	46.3	36.7	6.1	18.4	2:6 (0.3)	0.0	MEAIRFSNLSGLLILLMSLPLLVTSKDETVSCTMCSSCDNPCNPVPSSYSPPPPPPSSSGGGGSYYYSPPPPSSSGGVKYPPPYGGDGYGGYYPPPYYGNYGTPPPPNPIVPYFPFYYHTPPQGYSGSARLQNSLLFALFAVLLCFV
Class 10: EXT bias, AGP								
RRID_Locus_5314	47.6	53.5	47.6	15.0	5.9	2:1 (2.0)	0.0	MSSMRNKGGKKASLAFLFTTSALLVVIFSLCFPFLANAYRPPKSPTPGIYKPKSPPPKKPTYVPITPPTPGKPIKPRSPPPPYYQEDSAPSTVLPNLGKISPGVSKPGPPNKPPPKKKPPTDSQDSVSSTVLPNLGKKSPAVPKPLPPPRKKPPTLSQDSVPSTILPNLGKISPAVPKPLPPPRKKP
Class 11: EXT bias, high SP$_n$ (1)								
At1G44191.1	56.3	63.0	51.0	22.8	17.0	12:2 (6)	0.0	MKSLIFLVPILCLVISPTTTIASWPKPSEVSNEDMLMNTEHNHYFGNKLNFKDSKVWKCTYNNGSGAAVSISYPSPPHPPSPKPPTPSSRPPSPLSPKKSPPSPKPSPPPRTPKKSPPPPKPSSPPPIPKKSPPPPKPSSPPPTPKKSPPPPKPSSPPPSPKKSPPPPKPSPSPPKPSTPPPTPKKSPPSPPKPSSPPPSPKKSPPPPKPSPSPPKPSTPPPTPKKSPPPPKPSQPPPKPSPPRRKPSPPTPKPSTTPPSPKPSPPRPTPKKSPPPPTTPSPSYQDPWVHFANCISEFGPSAICKQQMEVSYRRGFLVSDYCCNLIVNMRHECTDVILGFFTDPFFVPLIRYTCHVKY
Class 12: EXT bias, high Tyr								
Solyc01g097680.2.1	48.0	70.9	59.2	21.6	31.0	6:26 (0.2)	0.0	MMESLGKLGQWSFLTVALAICLVASTVVADYSYEYTSPSPSHNSNKYYKSPSPSKYHVPTPYYKKPYPSHYYYKSPTPSKHAYYKSPSPAKYYKSHVPSKHYYKSPVATKYYYKFPTPSKHYYYKSPIVTKYYKSHVPSKHYYKSPVATKYYYKSPTPSKHYYKSPVPSKYYYKSPSPAKYYKSPTPSTNYYYKSPSPTKYYKSPAPSKYYKSPSPTKYYKSSPPPPKYYEQPPTYYNSPPPPYYEESTPSYKSSPPPPKTYEQSPTYYSPPPPPYYKETPTYASPPPPVKYEEPVTYASPPPPTFY
Solyc03g082790.2.1	47.1	71.9	55.0	30.6	19.4	0:20 (0)	0.0	MRRSLGNLGQLPLLAFALAICFVASTVVADYSYGYTSPSPTYTTKKYYKSPSPSPYYKKSEKHAEHSPSHYYKSPTPSKHYYKSPVVVKYYKSPAPSKKYYKSPSPSKYYYKSHTPSKKYYKSLSPAKYYKSPSPAKYYKSPTPSKHYYYKSPSPSKYYKSPSPAKYYKSPAPSKHYYYKSPSPAKYYKSPSPAKYYKSPAPSKHYYYKSPSPAKYYKSPSPAKYYKSPAPSKTITTSHHPL
Class 13: EXT bias, high PRP								
QCGM_Locus_1882	50.6	58.7	55.8	6.4	4.7	1:1 (1)	11.6	MELYVPLRGTFIFWKLVLFALSISFFVSLTSANYGHYPTPYTYVSPPPPSHGISPVYPPPKPPLKPPPVTKPPPVPKPPLLPPPVHKPPFHKPPPVHKPPPCPPPKYPPKPKPPIPKPPVYGPPLKPPPPPHKISPIYPPPIPKPKPSSILLQITTSTISQTSTYSKASLHP
C PRP Bias								
Class 3: PRPs								
Glyma15g23830.1	38.8	60.1	70.2	1.1	15.2	0:9 (0)	28.1	MSNMTSLSSLVLLLAALILFPQVLADYEKAPVYKPPIEKPPLYKPPIEKPPVYKPPVEKPPIYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVEKPPVEKPPVYKPPVEKPPVYKPPVENPPIYKPPVEKPPVYKPPIEKPPVYKPPVEKPPVEKPPVYKPPYGKPPYPKYPPIDDTHF
Class 15: PRP bias, AGP								
Medtr3g082100.1	45.9	47.3	55.9	18.3	2.1	1:1 (1)	4.4	MANYAIANVLILLLLNLSTLLNVLACPYCPYPSPKPPSHHPPIVKPPVHKPPKHSPTPKPPVHKPPRYPPKPSPCPPPSSTPKPPHHPKPPAVHPPHVPKPHPPYVPKPPIVKPPIVHPPYVPKPPVVKPPPYVPKPPVVRPPYVPKPPVVPVTPPYVPKPPIVFPPHVPLPPVVPSPPPYVPTPPIVKPPIVFPPHVPLPPVVPVTPPYVQPPPIVTPPTPTPPIVTPPTPPSETPCPPPPLVPYPPPPAQQTCSIDALKLGACVDVLGGLIHIGIGGSAKQTCCPLLQGLVDLDAAICLCTTIRLKLLNINLVIPLALQVLIDCGKTPPEGFKCPAS
Class 16: PRP bias, high EXT (SP$_n$ & Y)								
Selaginella_405021	50.0	64.1	66.1	11.8	57.8	22:30 (0.7)	0.0	MGPPRRTRMPMEATAALSLVVIALAFAAQVEASPYEYKSPPPPVYETPAPVYKYKSPPPPPPVYKYASPPPPVYHAPAPVYKYKSPPPPVYHYSSPPPPVYKYKSPPPPVYHAPAPVYNSPPPPVYKYKSPPPPVYHAPAPVYKYKSPPPPVYHYTSPPPPVYKYKSPPPPVYHAPAPVYKYKSPPPPVYHYSSPPPPVYKYKSPPPPVYHAPAPVYKYKSPPPPVYHYSSPPPPVYKYKSPPPPVYHAPAPVYKYKSPPPPVYHYSSPPPPVYEYKSPPPPVYHAPAPVYKYKSPPPPVYKYQSPPPPVYHSPAPVEDDNKCCRIFLEIRTRVAAVETELLHRKREM
Class 17: PRP bias, high SP$_n$								
GGJD_Locus_290	58.4	60.9	63.5	13.7	19.8	8:0 (0)	7.6	MGSQNFAGALILLNFGTLLTIALACPQCSNPTPPIFPPKHPSKPPPVVKPPYIPKPPYVPKPPVAPPPPLVPKSPPPPPPVPKSPPVPKSPPPAPKVVSPPPPVPCPPPPLVPKSPPPPAPAKSPPPPPTVPKSPPPVPKPPVSKPPPPPPKVVSPPPPVPCPPPPPPKASPPPKQPTCPIDTLKLGGCVDLLGLPH
Class 18: PRP Bias, high Tyr (1)								
At4G15160.1 PRP17	49.8	49.1	54.2	12.4	1.1	0:1 (0)	1.8	MALLHKQNISFVILLLLGLLAVSYACDCSDPPKPSPHPVKPPKHPAKPPKPPTVKPPTHTPKPPTVKPPPPYIPCPPPPYTPKPPTVKPPPPPYVKPPPPPTVKPPPPPYVKPPPPPTVKPPPPPTPYTPPPPTPYTPPPPTVKPPPPPVVTPPPPTPTPEAPCPPPPPTPYPPPPKPETCPIDALKLGACVDVLGGLIHIGLGKSYAKAKCCPLLDDLVGLDAAVCLCTTIRAKLLNIDLIIPIALEVLVDCGKTPPPRGFKCPTPLKRTPLLG
Glyma09g12252.1 SbPRP2	40.5	63.6	77.3	0.9	21.8	0:16 (0)	34.1	MASLSSLVLLLAALILSPQVLANYENPPVYKPPTEKPPVYKPPVEKPPVYKPPVENPPIYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPVYKPPVEKPPIYKPPVEKPPVYKPPYGKPPYPKYPPTDDTHF
Glyma15g38091.1	36.5	53.7	63.2	11.9	15.1	0:17 (0)	0	MQVWGKINSSFQTFSLFSFHWCGLAVFSMRKPSSLQSALLRFGVPLLLVITFCYATSTAAATTQVSGNEELPKTNLNGHDEEAKFGFFHHKPIFKKHIPIPVYKPVPKPFPVYKPIPKPVPVPVYKPIPKPVYKPIPKPVPVYKPIPKPVPVPVYKPIPKPVYKPIPKPVYKPIPKPVPVYKPIPKPVPVPVYKPIPKPVYKPIPKPVYKPIPKPVPVYKPIPKPVPVYKPIPKPVYKPIPKPVPVYKPIPKPVPIYKPIPIFKPIPKPVPIYKPIPIFKPVPKPVPIYKPIPIFKPVPTPIPFFKPIPKPVPIVKPIPIPVFKPIPKPFPIVKPIP
D Shared Bias								
Class 19: Shared Bias, AGP								
Cre10.g417600.t1.3	58.3	58.2	41.4	38.6	5.2	5:4 (1.3)	0.0	MGKSGRGFKWHLSLLAVVVMSSALKGSAQPGRPAKIATMPTAPPPPAPVAPATYPSYPLRPPPGPYGPPGPTYPPTYSPPIYSPPMYSPPIYSPPIYSPPIYTPPIYSPPVYPPVMPSPEPSPEESPAPSPEPESSPSPAVASPLPLPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPPPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPSPVPGPDDSPREYKCKPDVGLWNGKVIADIKLRPPHDDFEAVTLCEKECNTRDGECTGYHYEKDGYCTLWSGDVRFTRGVPSNIIACIFIGYKMPDPPSPEPPLVPLPPPPSPEPSPSPSPSPVPAPEASPNPSPSPSPITIPSPSPSPLPAPLPMPSPSPAPVPSPRPDPSPSDDSPRNYKCKDDVGLWNGDLIDRLKLRPPLNDVDAAARCRAECNKRDGVCAGYDYASDGYCKLWGGDVRYNRGSPGDIIACIFTGYKMPDPPSPEPPSPPPPPPPPSPLPPAPPPPSPPPPSPPPPSPPPPSPSPPSPSSSPSPSPSPSPSPKSSPSPSPSPKPSASPSSPGSGRSYSCKNGLSLWNGNQLEVITIKLPTNDEAVAKCQENCNNKPGCTGFHYKTTTRCYLWGGANLLFTKPYPSDVMACIWTQQFSRSSSNR
Cre02.g089400.t1.3	48.3	48.4	31.6	31.0	0.0	0:0 (0)	0.0	MRPGLHVAALLVAAITSAVASPDITCTGTIFVKSGYSGASYDITLTPQGYSDRSYVWDLKDVPSGVPGKPTWDDLANSLKLSCTCLDDSTASCAADYEKIIMRVWAEPGSRGQGGGDSLDVTCAASADGKSCAGALSTMPAGWGSRLSAVAILYNNAGYTLSKSRPETPITPTGGVLKRLLQSDDDDDDDDDDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDDDGDDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDDDDDDGDDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDDGNDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDDGNDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDGNDDDDDRSPSPKPSPSPQPSPSPKPSPSPKPSPSPRPSPSPDNDDDDDDDDGNDDDDDRSPSPKPSPSPQPSPSPKPSPSSKPSPSPRPSPSPDNDDDDDNDDDRSPSPKPSPSPKPSPSRPPPGSGRRRPNGRPRPARRVNGRNGRRLLL
Class 20: Shared Bias, high EXT (SP$_n$ & Y) (7)								
At1G21310.1 EXT3	49.7	74.2	73.1	1.9	74.7	41:43 (1)	0.0	MGSPMASLVATLLVLTISLTFVSQSTANYFYSSPPPPVKHYTPPVKHYSPPPVYHSPPPPKKHYEYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKEKYVYKSPPPPPVHHYSPPHHPYLYKSPPPPVHY
At5G11990.1 EXT38	51.4	49.7	36.5	13.3	17.1	5:3 (1.7)	0.0	MGLIVIIIALVMVMVVASSPSDQTNVLTPLCISECSTCPTICSPPPSKPSPSMSPPPSPSLPLSSSPPPPPPHKHSPPPLSQSLSPPPLITVIHPPPPRFYYFESTPPPPPLSPDGKGSPPSVPSSPPSPKGQSQGQQQPPYPPPYFYFYTASNATLLFSSSFLIALVVSSLSIFLNGSLV
At5G19800.1 EXT39	47.9	52.1	45.8	3.1	21.9	3:1* (3)	0.0	MVSQRFSLTLVFLLAILITVANGNDNRKLLTSYKYSPPPPPVYSPPISPPPPPPPPPPQSHAAAYKRYSPPPPPSKYGRVYPPPPPGKSLWSFLNL

Figure 4. *(Figure continues on following page.)*



	Bias (%)			Motifs				Sequence
	PAST	PSKY	PVYK	%AGP	%EXT	SPn:Y (ratio)	%PRP	
Class 21 Shared Bias, high SPn (2)								
At5G19810.1 EXT19	57.8	57.4	55.4	3.2	41.4	17:2 (8.5)	0.0	MVSRRFSLILLFLLATLITVANGNNNRKLLTSAPEPAPLVDLSPPPPPVNISSPPPPVNLSPPPPPVNLSPPPPPVNLSPPPPPVNLS PPPPPVLLSPPPPPVNLSPPPPPVNLSPPPPPVLLSPPPPPVLLSPPPPPVNLSPPPPPVLLSPPPPPVLFSPPPPTVTRPPPPPTIT RSPPPPRPQAAAYYKKTPPPPPYKYGRVYPPPPPPPPQAARSYKRSPPPPPPSKYGRVYSPPPPGKSWLWFLKL
At3G19020.1 PEX1	45.3	45.3	39.4	11.7	14.4	25:3 (8.3)	0.0	MTRRTMEKPFGCFLLLFCFTISIFFYSAAALTDEEASFLTRRQLLALSENGDLPDDIEYEVDLDLKFANNRLKRAYIALQAWKKAFYS DPFNTAANWVGPDVCSYKGVFCAPALDDPSVLVVAGIDLNHADIAGYLPPELGLLTDVALFHVNSNRFCGVIPKSLSKLTLMYEFDVS NNRFVGPFPTVALSWPSLKFLDIRYNDFEGKLPPEIFDKDLDAIFLNNNRFESTIPETIGKSTASVVTFAHNKFSGCIPKTIGQMKNL NEIVFIGNNLSGCLPNEIGSLNNVTVFDASSNGFVGSLPSTLSGLANVEQMDFSYNKFTGFVTDNICKLPKLSNFTFSYNFFNGEAQS CVPGSSQEKQFDDTSNCLQNRPNQKSAKECLPVVSRPVDCSKDKCAGGGGGGSNPSPKPTPTPKAPEPKKEINPPNLEEPSKPKPEES PKPQQPSPKPETPSHEPSNPKEPKPESPKQESPKTEQPKPKPESPKQESPKQEAPKPEQPKPKPESPKQESSKQEPPKPEESPKPEPP KPEESPKPQPPKQETPKPEESPKPQPPKQETPKPEESPKPQPPKQETPKPEESPKPQPPKQEQPPKTEAPKMGSPPLESPVPNDPYDA SPIKKRRPQPPSPSTEETKTTSPQSPPVHSPPPPPPVHSPPPPVFSPPPPMHSPPPPVYSPPPPVHSPPPPPVHSPPPPVHSPPPPVH SPPPPVHSPPPPVHSPPPPVHSPPPPVQSPPPPPVFSPPPPAPIYSPPPPPVHSPPPPVHSPPPPPVHSPPPPVHSPPPPVHSPPPPPV HSPPPPVHSPPPPSPIYSPPPPVFSPPPKPVTPLPPATSPMANAPTPSSSESGEISTPVQAPTPDSEDIEAPSDSNHSPVFKSSPAPS PDSEPEVEAPVPSSEPEVEAPKQSEATPSSSPPSSNPSPDVTAPPSEDNDDGDNFILPPNIGHQYASPPPPMFPGY
Class 22: Shared Bias, high Tyr								
Potri.010G072200.1	34.6	45.8	44.1	1.1	14.0	0:9 (0)	0.0	MQITSLLVLFVGVVVLATPSFADYYKPPKYEPKPPYFKPPKVEKPFPEHKPPVYKPPKIEKPPVYKPPPVYKPPIEKPPVYKPPPVYK PPIEKPPVYKPPPSTSHHQSTSHQLRSHQSTSHQLRSHQSTSLQRLRSHHPFTSHCHLMVTIRDTLQLKMRNISSQKTELNFLYIDMQ EYY
Class 23: Shared Bias, high PRP (1)								
At2G27380.1 PRP6	61.0	66.1	66.6	23.7	0.4	0:1 (0)	27.1	MRVPLIDFLRFLVLILSLSGASVAADATVKQNFNKYETDSGHAHPPPIYGAPPSYTTPPPPIYSPPIYPPPIQKPPTYSPPIYPPPIQ KPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPSYSPPVKPPPVQMPPTPTYSPPIKPPPVHKPP TPTYSPPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHKPPTPIYSPPIKPPPVHKPPTPIY SPPVKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPPVKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPV KPPPVHKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPIKLPPVKPPTPIYSPPVKPPP VHKPPTPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTPTYSPPVQPPPVQKPPTPTYSPPVKPPPIQKPPTPTYSPPIKPPPVKPP TPTYSPPIKPPPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPPTPT YSPPIKPPPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPPTPTYSPPVKPPPVQVPPTPTYSP PVKPPPVQVPPTPTYSPPIKPPPVQVPPTPTTPSPPQGGYGTPPPYAYLSHPIDIRN
E <15% HRGP motifs								
Class 24: <15% motif (7)								
At1G30795.1	45.9	33.9	30.3	8.3	3.7	1:0 (0)	0.0	MASSTHSYFTTLALTLILIFRLIPETTASRHLNGKNPAVIGVTTTSEKYIVPTPLPPFLRPFFPPLQFAAAPFGGNIPQPPLPSPPPT FLPCLPGFKFPPPFQSRKPTPP
At2G22510.1	54.0	37.1	27.4	4.8	0.0	0:0 (0)	4.0	MAHWLSSLVIALTFTSFFTGLSASRHLLQSTPAITPPVTTTFPPLPTTTMPPFPPSTSLPQPTAFPPLPSSQIPSLPNPAQPINIPNF PQINIPNFPISIPNNFPPFNLPTSIPTIPFFTPPPSK
At5G09530.1 PRP10	35.1	46.5	49.5	12.4	0.0	0:0 (0)	0.0	MALMKKSLSAALLSSPLLIICLIALLADPFSVGARRLLEDPKPEIPKLPELPKFEVPKLPEFPKPELPKLPEFPKPELPKIPEIPKPE LPKVPEIPKPEETKLPDIPKLELPKFPEIPKPELPKMPEIPKPELPKVPEIQKPELPKMPEIPKPELPKFPEIPKPDLPKFPENSKPE VPKLMETEKPEAPKVPEIPKPELPKLPEVPKLEAPKVPEIQKPELPKMPELPKMPEIQKPELPKLPEVPKLEAPKVPEIQKPELPKMP ELPKMPEIQKPELPKMPEIQKPELPKVPEVPKPELPTVPEVPKSEAPKFPEIPKPELPKIPEVPKPELPKVPEITKPAVPEIPKPELP TMPQLPKLPEFPKVPGTP
At5G59170.1 PRP12	46.9	68.4	67.7	4.5	6.3	0:6 (0)	0.0	MKSCLATFALAVLVTQGSIFSLCLASSASTNSVPNWFHHWPPFKWGPKFPYSPPKPPPIEKYPPPVQYPPPIKKYPPPPYEHPPVKYP PPIKTYPHPPVKYPPPEQYPPPIKKYPPPEQYPPPIKKYPPPEQYSPPFKKYPPPEQYPPPIKKYPPPEHYPPPIKKYPPQEQYPPPI KKYPPPEKYPPPIKKYPPPEQYPPPIKKYPPPIKKYPPPEEYPPPIKTYPHPPVKYPPPPYKTYPHPPIKTYPPPKECPPPPEHYPWP PKKKYPPPVEYPSPPYKKYPPADP

Figure 4. Illustration of the parameters used for MAAB classification of HRGPs. Where possible, for the Arabidopsis sequences, we have included both the gene names and the nomenclature designated by Showalter et al. (2010; in blue text). The total number of Arabidopsis sequences identified for a given class is shown in parentheses, and up to four examples are shown. If no Arabidopsis sequence was present in a given class, then sequences from other species, either from Phytozome or 1KP, were used. For class 18, an Arabidopsis sequence (At4G15160.1) does not have the expected features of this class due to the partially order-dependent assignment of the minor classes (see "Materials and Methods"). The columns reporting amino acid bias, as used to classify sequences into AGP bias (orange), EXT bias (red), PRP bias (purple), or shared bias (yellow), are shaded as for Figure 3. Shading of motifs is used to highlight the number of hybrid sequences that satisfy 10% or greater motifs for any given HRGP class. SP_n:Y is reported as the number of SP_n motifs:number of Y motifs (ratio of SP_n:Y reported as a fraction). White text for the SP_n:Y ratio indicates that the sequence does not satisfy at least one of the criteria for CL-EXT: at least two SP_n and two Y motifs (indicated by asterisks) or a ratio of SP_n:Y between 0.25 and 4 (reported here as 0 if either value is 0). The order of motif searching is CL-EXT motifs first, followed by PRP motifs, and, finally, AGP motifs. Sequences are shown with HRGP motifs (as used for classification) highlighted as follows: light blue for AGP motifs, red for EXT SP_{3-5} motifs, dark blue for Y-based EXT motifs, and olive green for PRP motifs. In cases where motifs overlap, as occurs frequently in the shared bias classes (18–23), shading shows only the accepted, first identified, motif.

combinations of HRGP motifs. For example, in class 5 (AGP bias, high EXT [SP_n and Y]), EXT30 (At1G02405) has few AG-glycomotifs and more EXT motifs with at least two SP_n and two Y motifs (Fig. 4A). The EXT bias classes (Fig. 4B) include the classical CL-EXTs (class 2) that have a similar number of SP_n to Y motifs (ratio between 0.25 and 4), reflecting the predominantly alternating nature of these motifs in CL-EXTs. In contrast, the minor EXT bias classes 9 to 13 include the rare GPI-anchored EXTs (class 9) and EXTs with an overrepresentation of AGP motifs (class 10), SP_n motifs (class 11, SP_n:Tyr > 4), Y motifs (class 12, SP_n:Tyr < 0.25), or PRP motifs (class 13). Sequences such as EXT39 (At5G19800) with EXT bias but only one SP_n and/or one Y motif cannot be typical CL-EXTs with predominantly alternating SP_n and Y motifs and, therefore, were included in class 20 (shared bias, high EXT [SP_n and Y]) to flag that they are unusual (Fig. 4D). Also within class 20, EXT3 (At1G21310) has features of a classical class 2 CL-EXT yet is classified in class 20 due to a shared bias (less than 2% difference [Δ]) between percentage PSKY (74.2%) and percentage PVYK (73.1%). This highlights the ability of MAAB to identify outliers from the classical CL-EXTs and is an effective method to distinguish EXTs with different SP_n/Y motifs. In order to classify AtEXT3 as a class 2 CL-EXT, the shared bias parameter value could be changed from 2% to 1%. We did not adjust MAAB, as we wanted to more readily observe HRGP variation; whether our classification reflects functional differences is as yet untested. Given the role of Y motifs in intermolecular and intramolecular cross-linking (Held et al., 2004; Lamport et al., 2011), both the arrangement and composition of SP_n/Y motifs are likely to impact function. Therefore, the identification and characterization of the diversity of EXT motifs could guide important functional studies.

The PRPs are classified in a similar way (Fig. 4C), with classical (non-chimeric) PRPs in class 3 and other sequences with PRP bias in classes 14 to 18 (Fig. 3), where class 14 (GPI-anchored PRP) is allowed for even though they have never been described. Very few PRP sequences were identified, and this is likely to be because almost all PRPs characterized to date are chimeric and, therefore, excluded in stage 1. Another possibility is that more PRP motifs need to be incorporated into the MAAB pipeline. It is likely that such motifs have been identified in class 24 (see discussion below), and with validation of glycosylation on novel Pro-rich motifs, MAAB could be adapted to improve the recovery of PRPs. Sequences with no clear amino acid bias (Δ amino acid bias < 2) are put into the shared bias category, and within most of these classes (classes 19–23), there is a diversity of HRGP motifs because many of these sequences are hybrid HRGPs. For example, PEX1 (At3G19020) in class 21 has exactly the same percentage amino acid bias for both AGP and EXT. Motif shading within the sequence clearly shows that it contains both AGP and EXT motifs and is a chimeric HRGP (Fig. 4D), because of its relatively poor match to PFAM domain PF08263 (data not shown). In the shared bias category, there is not a strict correlation between the class and motif type because the classes are assigned in a predetermined (fixed) order (see "Materials and Methods"). The final class, MAAB class 24, contains sequences with less than 15% known HRGP motifs (Fig. 4E) and will be discussed in more detail below.

Evaluation of the MAAB Pipeline Using Phytozome Proteomes

Additional analyses performed on the Arabidopsis data set downloaded from Phytozome provided confidence that the MAAB pipeline is robust and effective at classifying GPI-AGPs, CL-EXTs, and hybrid HRGPs into consistent classes based on their unique features (Supplemental Table S2). In Arabidopsis, 17 GPI-AGPs were identified by MAAB, including a new GPI-AGP (At3G27416.1; hereafter AtAGP59), as compared with the 16 classical AGPs identified by BIO OHIO/manual curation (Showalter et al., 2010; Table I; Supplemental Table S2; Supplemental Fig. S3). Of the 59 EXTs identified by BIO OHIO/manual curation (Showalter et al., 2010), many do not meet our stringent criteria, as they are either chimeric (e.g. PERK, HAE, and LRX) and/or have lower amino acid bias (less than 45% PSKY), no ER signal, no cross-linking motifs, or they have a shared bias with PRPs (classes 20 and 21; Fig. 4; Supplemental Table S2). This resulted in MAAB finding 16 CL-EXTs in Arabidopsis (Table I), one of which is a chimeric Leu-rich repeat/EXT (LRX1; At1G12040) and was not excluded by the MAAB pipeline because of its relatively poor match to PFAM domain PF01816 (data not shown). No new CL-EXTs were identified, and no obvious false positives were found among either MAAB class 1 or 2 sequences.

Of the 27 classical and Lys-rich AGPs identified in *P. trichocarpa* by BIO OHIO 2.0/manual curation (Showalter et al., 2016), MAAB classified 10 as class 1, four as class 4, and eliminated eight due to greater than 95% identity with another sequence (stage 1a), low percentage PAST (stage 1b), or containing a lipid-transfer protein/hydrophobin PFAM domain (stage 1e; Supplemental Table S3). Similarly, many of the EXTs identified by BIO OHIO 2.0/manual curation (Showalter et al., 2016) were either reclassified into other HRGP classes or eliminated by MAAB (Table I; Supplemental Table S3). Five of the 22 *P. trichocarpa* short EXTs were classified as non-GPI-AGPs because none satisfied the requirement for at least two SP_n and two Y motifs, and two also had more AGP motifs than EXT motifs. Six other short EXTs were classified in other MAAB classes (5, 20, 21, or 24), and 11 others were eliminated at stage 1 of the MAAB classifier (either stage 1b, 45% or greater bias [10], or stage 1c, 90 or greater amino acids [1]; Table I; Supplemental Table S3).

A noteworthy outcome of the MAAB pipeline was the low number of PRP sequences, including their apparent absence from Arabidopsis (Table I). All the Arabidopsis and *P. trichocarpa* PRPs identified by BIO OHIO/manual curation (Showalter et al., 2010, 2016) were excluded by MAAB, mostly due to being chimeric, but others were excluded due to the absence of an ER signal peptide or low amino acid bias (Supplemental Tables S2 and S3, respectively). Only a single PRP was identified in soybean (Table I; Fig. 4), one of the few species in which non-chimeric PRP gene sequences have been reported (Datta et al., 1989). The previously cloned SbPRP1 (Glyma09g12198.1), SbPRP2 (Glyma09g12252.1; Fig. 4), and SbPRP3 (Glyma11g21616.1; Hong et al., 1987, 1990) are found in HRGP class 18 (PRP bias, high Y) because they have a higher number of EXT motifs than PRP motifs. A sequence (Glyma10g33050.2) similar to a fourth soybean PRP (ENOD2; Franssen et al., 1987) has only 9.8% HRGP motifs and, therefore, is found in MAAB class 24. These findings highlight the need to increase the range of PRP motifs used by MAAB. Only one Arabidopsis PRP identified by BIO OHIO/manual curation (Showalter et al., 2010) was classified as an HRGP by MAAB (At2G27380, PRP6, class 23), and two (At5G09530.1, PRP10 and At5G59170.1, PRP12) are in class 24 (less than 15% known motifs; Fig. 4, D and E; Supplemental Table S2). The remaining nine of 12 Arabidopsis PRPs identified by BIO OHIO/manual curation (Showalter et al., 2010) were eliminated at stage 1: four due to low bias (stage 1b), four due to the presence of PFAM domains (stage 1e), and one had no ER signal sequence (stage 1f). In contrast, none of the 16 *P. trichocarpa* PRPs identified by BIO OHIO 2.0/manual curation (Showalter et al., 2016) were classified as HRGPs by MAAB, with only two being retained by MAAB (class 24; less than 15% known motifs); two were eliminated due to low bias (stage 1b), and the remaining 12 had a PFAM domain as noted by the authors (Supplemental Table S3). MAAB input parameters were not adjusted to detect additional non-chimeric PRPs due to the pipeline's ability to detect and classify class 1, 2, and 4 HRGPs with high reliability in the broad taxonomic test of Phytozome proteomes.

MAAB identified GPI-AGPs (class 1) and non-GPI-AGPs (class 4) in all 15 proteomes (Table I). In addition to Arabidopsis, genes encoding AGPs have been found previously in rice (*Oryza sativa*; Ma and Zhao, 2010) and the moss *Physcomitrella patens* (Lee et al., 2005). The MAAB pipeline identified 11 GPI-AGPs in rice compared with eight in a previous study (Ma and Zhao,

2010) from similar data sets (Rice Genome Annotation Project, release 5). Differences are likely due to different percentage PAST thresholds (45% compared with 50%; data not shown). The report of AGPs from *P. patens* was not a detailed bioinformatics study but rather used a Hyp-containing protein backbone sequence to identify ESTs using tBLASTn (Lee et al., 2005). The two GPI-AGPs found in that study, PpAGP1 and PpAGP2, were both identified by the MAAB pipeline (Pp1s143_12V6.1 and Pp1s338_8V6.1, respectively; Supplemental Fig. S3).

An intriguing finding was the absence of CL-EXTs in the two volvocine algal genomes (Table I). The five volvocine algal HRGP class 5 and 6 sequences (e.g. Cre16.g693200.t1.3) all have SP_n glycomotifs in a domain separate from one or more domains with scattered Y residues, as observed previously for algal HRGPs (Fig. 1, I–K). The ability of the MAAB pipeline to capture diverse HRGPs from a wide range of organisms with high accuracy proved its utility and, hence, suitability to analyze the larger 1KP transcriptome data.

Using the MAAB Pipeline on 1KP Transcriptomic Data Sets: Multiple *k* Assembly Improves HRGP Detection

The 1KP Initiative has generated large-scale transcript sequence data for a diverse range of species in the green plant lineage (www.onekp.com). Investigation of RNA sequencing samples from a wide range of multicellular species poses many challenges, including the diversity and type of tissues sampled and potentially large numbers of partial sequences. Reconstruction of transcripts from short-read sequences presents issues for De Bruijn graph transcriptome assemblers, in particular the choice of *k*-mer size, which must be specified a priori (Nagarajan and Pop, 2013; Yang and Smith, 2013; Rana et al., 2016). The *k*-mers are subsets of contiguous, overlapping nucleotides of a defined length (*k*) that are generated during the assembly of short-read sequences. Our initial investigations of the protein predictions provided by the 1KP Consortium, which employed 25-mers (Xie et al., 2014), resulted in limited recovery of HRGP sequences, particularly the repetitive CL-EXTs and PRPs (Johnson et al., 2017).

To improve the assembly of HRGP genes, we reassembled sequence reads for all samples using four *k*-mer sizes (39, 49, 59, and 69) to ensure that *k*-mers were large enough to span putative glycomotif repeats (Data File 2). The full output of combined (multiple) *k*-mer results (compared with $k = 25$) is reported in the companion article (Johnson et al., 2017). The MAAB-designated HRGPs captured by the four *k*-mer size assemblies greatly enhanced the recovery of the major classes of HRGPs across all plant species. Each individual *k*-mer yielded a significant number of discrete HRGP sequences as well as many sequences detected by more than one *k*-mer (Supplemental Fig. S4). A potential problem with this approach is repeated detection of the same sequences. This has been alleviated somewhat by the identification and removal of excess copies of substantially identical (95% or greater) sequences within MAAB (stage 1, step 1a; Fig. 3; see "Materials and Methods").

Comparison of the four *k*-mer assemblies showed that no single assembly parameter was best for any given HRGP class (1–4; Supplemental Fig. S4). For example, of the 5,754 total GPI-AGPs identified in the 1KP transcriptomes, 1,630 GPI-AGPs were identified only with $k = 39$, another 623 with $k = 49$, 284 with $k = 59$, and 159 with $k = 69$. The remaining 3,058 GPI-AGPs were identified with two, three, or four different *k*-mers (Supplemental Fig. S4A). The *k*-mer(s) that produced each sequence is reported in the MAAB output (Data Files 3 and 4).

Comparison of 1KP HRGPs Identified by the MAAB Pipeline to Experimentally Confirm Hyp-Containing Peptides

HRGPs have been isolated from a diverse range of plant species and subjected to protein/peptide sequencing (Supplemental Table S1). To further validate the effectiveness of the MAAB pipeline, experimentally confirmed Hyp-containing peptides identified in the literature (Supplemental Table S1) were used in BLAST searches against the HRGP proteins identified in the 1KP multiple assemblies. Peptides associated with PRPs, CL-EXTs, and AGPs were largely found in the expected MAAB classes (Supplemental Table S1). For example, four peptides from a Douglas fir (*Pseudotsuga menziesii*) EXT (Fong et al., 1992; Kieliszewski et al., 1992) matched a CL-EXT sequence (AREG_Locus_407) in the conifer *Nothotsuga longibracteata* data set (Supplemental Table S1). These data support our conclusion that authentic HRGPs are identified by the MAAB pipeline. Only a few Hyp-containing peptides have been experimentally confirmed in volvocine algae and bryophyte species, and these are associated with AGP motifs (Supplemental Table S1). Although a match for the Hyp-containing peptide from *Volvox carteri* was not found in the 1KP output, our findings show that AGPs are common in chlorophyte and streptophye algae (Johnson et al., 2017).

MAAB Class 24: A Resource to Mine for New IDPs and PRP Motifs

Class 24, the class with less than 15% known HRGP motifs, was the most abundant class overall and accounts for 46.8% (39,933 of 85,237) of sequences retained by the MAAB pipeline (Johnson et al., 2017). In general, the largest proportion of class 24 sequences was found in the algal clades compared with land plants (embryophytes). The low proportion of common HRGP motifs XP_1, XP_3, and XP_4 (where X = A, T, S, Y, K, V, or G) in class 24 sequences is demonstrated graphically in Supplemental Figure S5. No other XP motifs were noticeably prominent (for X = F, H, and L; for all other X [data not shown]; Supplemental Fig. S5). Preliminary analysis suggests that there is a large diversity of sequences in MAAB class 24 (see Data Files 3 and 4). This class was checked for Pro-rich motifs from mammals and other organisms, and no motifs were found to be well represented. Some plant sequences similar to rice OsRePRP1 (Tseng et al., 2013) and AtPRP10 (Rashid and Deyholos, 2011) were found, with a total of 7.3% of class 24 proteins containing either an OsRePRP1 (0.5%) or an AtPRP10/AtPELPK1 (6.8%) Pro-rich motif (Table II).

The amino acid bias of AtPRP10/AtPELPK1 is distinct from that of other HRGPs and contains the repeated motif PE[L/I/V]PK (Fig. 4E). AtPELPK1 is localized to the cell wall (Rashid, 2014) and is an IDP, yet it remains uncertain if it and OsRePRP1 (Tseng et al., 2013) are bona fide Hyp-containing glycoproteins. Determining if and where Pro-to-Hyp modification (and subsequent glycosylation) occurs in the novel motifs identified within class 24 will require experimental approaches such as in planta expression. Recombinant proteins, for example, can be used to test the Hyp-contiguity hypothesis in the diversity of new sequence contexts revealed by 1KP (Shpak et al., 2001; Tan et al., 2003; Estévez et al., 2006). Further analysis of class 24 sequences also may uncover other novel HRGP motifs and/or new IDP proteins that are not HRGPs.

Versatility of the MAAB Pipeline for HRGPs

To optimize the recovery of specific classes of HRGPs, the MAAB pipeline can be readily adapted. For example, the bias threshold could be decreased to 1% or 0.5% to increase the number of sequences in the major classes (1–4). The pipeline can be modified to classify chimeric HRGPs by masking PFAM domains and applying the classifier to the unmasked portion of sequences. Another feature that is able to be modified is the order of motif counting, as it becomes important where there is partial overlap of motifs between two HRGP classes, such as the VYK motif in CL-EXT and the PPVYK motif of PRPs (see "Materials and Methods"). In this implementation of the MAAB pipeline, EXT motifs were counted first, then PRP motifs. This contributed to some candidate PRPs having a reduced PRP motif count and, therefore, not being identified as PRPs. For example, the 1KP sequence TJMB_Locus_208 from *Glycine soja* is placed into class 18 (PRP bias, high Y) due to having an equal count of EXT and PRP motifs (six EXT motifs, all VYK, as part of the longer PRP motif PPVYK and six PRP motifs, PPVEK; Data File 3). If the PRP motifs were counted first, the result would be 12 PRP motifs and no EXT motifs.

New motifs also can be added to the MAAB pipeline, such as repeats identified using tandem repeat annotation

Table II. *Occurrence of known Pro-rich motifs in class 24 proteins (less than 15% HRGP motifs)*

Motifs	Polymer	Vascular Plants[a]	Non-vascular Plants[b]	References
Motifs identified				
PEPK	OsRePRP1	165	22	Tseng et al. (2013)
PEPKPKPEPK	OsRePRP1	17	0	Tseng et al. (2013)
PELPK	AtPRP10/AtPELPK1	17	4	Rashid and Deyholos (2011
PEXPK	AtPRP10/AtPELPK1	1,563	776	Rashid and Deyholos (2011)
(PEXPK)$_2$	AtPRP10/AtPELPK1	346	4	Rashid and Deyholos (2011)
KPPP	120-kD Douglas fir extensin[c]	855	309	Fong et al. (1992); Schultz et al. (1997)
PGQGQQ	Gluten, wheat	0	1	Roberts et al. (2015)
PPPVHL	γ-Zein	1	1	Matsushima et al. (2008)
(PPG)$_2$	Collagen	1	8	Matsushima et al. (2008)
(VPGXG)$_1$	Elastin, human	266	639	Roberts et al. (2015)
(VPGXG)$_2$	Elastin, human	2	2	Roberts et al. (2015)
PGMG	Biomaterialization molecule, sea urchin	8	16	Matsushima et al. (2008)
Motifs not identified				
GYPPQQ	Synexin, *Dictyostelium*	0	0	Matsushima et al. (2008)
PFPQQPQQ	ω-Gliadin	0	0	Matsushima et al. (2008)
AKPSYPPTYK	Mussel adhesive protein	0	0	Matsushima et al. (2008)
VTSAPDTRPAPGSTAPPAHG	MUC1	0	0	Matsushima et al. (2009)
APDTRPA	MUC1 epitope	0	0	Pepbank[d]
GYYPTSPQQ	Gluten, wheat	0	0	Roberts et al. (2015)
PQGPPQQGGW	Acidic PRP, human saliva	0	0	Bennick (1987)
PQGPPPQGG	Basic PRP, human saliva	0	0	Bennick (1987)
KPEGPPPQGGNQSQGPPPPPG	Human PRB4 salivary gland PRP	0	0	Matsushima et al. (2009)
PPPPGGPQPRPPQG	Human PRB4 salivary gland PRP	0	0	Matsushima et al. (2009)

[a] Data File 3, eudicots to monilophytes. [b] Data File 4, lycophytes to algae. [c] In both plant examples, two of three Pro residues in KPPP are hydroxylated to KPOO (Supplemental Table S1). [d] http://pepbank.mgh.harvard.edu/interactions/details/16312.

libraries (Johnson et al., 2017), and, with an iterative approach, would allow refinement to suit specific data sets/sequences of interest. We also have identified additional features, such as a bias and positioning of specific amino acids like Lys, Met, Gln, and Asp/Glu, in particular GPI-AGPs (Johnson et al., 2017). These features also could be incorporated into the MAAB classification process.

The MAAB pipeline was built on existing knowledge of HRGP motifs but allows scope to find new motifs by providing categories that reflect different amino acid biases and functional motifs. A major technical challenge is the difficulty of predicting the PTMs of HRGPs, including the sites of Pro hydroxylation and types of glycosylation. This will require detailed structural analysis of many members of each multigene family, from key transitions throughout the plant lineage. Even within a single species, glycosylation is tissue dependent and directed by the context of amino acids surrounding the glycomotif (Tan et al., 2003; Shimizu et al., 2005; Kurotani and Sakurai, 2015). As we gain further knowledge of the PTMs on HRGPs, these can be incorporated into the MAAB pipeline. For example, further features can be added to particular orders or families, such as differences in glycosylation that occur in algal HRGPs compared with bryophytes (Johnson et al., 2017).

CONCLUSION

The MAAB Pipeline: A New Bioinformatics Resource to Study Intrinsically Disordered Proteins

We have developed and demonstrated the versatility and utility of an open-access pipeline for identifying and classifying the HRGP superfamily of IDPs by MAAB that is stringent, consistent, and flexible. The utility of MAAB reaches far beyond HRGPs. Features of other IDPs can be incorporated into MAAB and used to identify IDPs in available genomes/transcriptomes. This would allow the identification of, for example, the repetitive domains of elastins (Roberts et al., 2015), salivary PRPs (Manconi et al., 2016), insect silks (Starrett et al., 2012), AGL proteins from arbuscular mycorrhizal fungi (Schultz and Harrison, 2008), and LATE EMBRYOGENESIS ABUNDANT proteins in plants (Sun et al., 2013), with low rates of false negatives and false positives.

With increasing identification and knowledge of IDPs, the importance of these molecules in different biological contexts is becoming apparent. Since IDPs typically contain motifs that mediate multiple molecular interactions, they are commonly associated with signaling pathways and have recently been linked to a number of human diseases (Babu, 2016). This is emphasized by a bioinformatics study of transcription factors that showed that extended regions of disorder are common in the regulatory domains (Liu et al., 2006). Study of individual IDPs over evolutionary time scales presents an exciting opportunity to investigate their functional significance in different biological contexts.

Tracking IDPs in Large Data Sets and over Evolutionary Time Scales

Implementation of the MAAB pipeline for the HRGP superfamily of IDPs has highlighted both the enormous strengths and also the limitations of working with large transcriptomic data sets and provides strategies to optimize the recovery of IDP sequences, a necessary first step to the identification of putative orthologs. Our study highlights the importance of a multiple-*k*-mer assembly approach for the recovery of HRGPs and is a strategy that should be employed when searching for IDPs. This is particularly relevant for transcriptomes based on short-read sequencing data. In the future, this issue will be reduced with the development of more reliable long-read sequencing technologies.

Tools such as MAAB provide the basis for further approaches to track individual IDPs throughout evolution. This is not without difficulty, as the functional motifs in IDPs are small and clustered and can be rapidly gained and lost during evolution (Forman-Kay and Mittag, 2013). This can result in sequences with variable numbers of repeat motifs and diverse sequence lengths that cannot be aligned in a meaningful way with the tools developed for folded proteins. In our companion article (Johnson et al., 2017), we provide strategies to track specific plant IDPs throughout evolution using the MAAB output from the 1KP data and provide a platform for further investigation of HRGPs.

MATERIALS AND METHODS

Analysis of Plant Genome Data

Protein data from the completed genomes of the 15 species listed in Table I as well as *Ostreococcus lucimarinus* were downloaded from Phytozome version 9 (https://phytozome.jgi.doe.gov/pz/portal.html).

Sequence Analysis of Arabidopsis AGPs and EXTs

Arabidopsis (*Arabidopsis thaliana*) AGPs and EXTs (untrimmed sequences) are as reported by Showalter et al. (2010) with sequences downloaded from TAIR version 10 (https://www.arabidopsis.org; Supplemental Fig. S6). Multiple sequence alignments were performed using MUSCLE (Edgar, 2004) in MEGA 6.06 (Tamura et al., 2013) and the default settings, as recommended (Hall, 2013). Pairwise identity matrices were generated from Arabidopsis alignments by importing sequence alignments into Geneious 8.1.3 (Biomatters; Supplemental Fig. S1). Shading of HRGP motifs in the aligned sequences was done manually. IDP analysis was performed at http://www.pondr.com using three different predictors: VL-XT (red; Romero et al., 1997, 2001; Li et al., 1999), VSL2 (green; Obradovic et al., 2005), and VL3 (blue; Radivojac et al., 2003).

Phylogenetic Analysis

Phylogenetic analyses were performed on a desktop computer using MEGA 6.06 (Tamura et al., 2013). Sequences were first aligned using MUSCLE (Edgar, 2004) using the default settings, and no trimming of sequences was performed.

Data Sets

A summary of data sets and methods is also available at http://services.plantcell.unimelb.edu.au/hrgp/index.html.

Data analysis was based on 1,282 samples downloaded from the official 1KP mirror (onekp.westgrid.ca; as of March 2014). No compensation was performed for partial sequences (e.g. using targeted assembly methods or scaffolding using genomic resources). Preliminary analyses (Johnson et al., 2017) used the 1KP Consortium's *k*-mer = 25 assembly (Xie et al., 2014), whereas all subsequent analyses used the multiple *k*-mer data generated as summarized here. Sample read sets were assembled with Oases (Schulz et al., 2012) using four different *k*-mers (39, 49, 59, and 69) and open reading frames identified using getorf from the EMBOSS toolkit (http://emboss.sourceforge.net/). The predicted proteins were subsequently screened with an in-house BioPerl script to identify compositionally biased likely HRGP family members (Fig. 3). This preliminary screen identified 3,590,006 sequences (across all four assembly *k*-mers) for input to the MAAB pipeline (Data File 2). The screening script is available at http://services.plantcell.unimelb.edu.au/hrgp/index.html.

Removal of Contaminated Data Sets and Additional Data Sets

The integrity of all 1KP data sets was checked by rRNA sequencing and a list of contaminated data sets provided by the 1KP Consortium (https://pods.iplantcollaborative.org/wiki/display/iptol/Sample+source+and+purity). Contaminated data sets were removed after MAAB analysis and are not included in the output files unless noted otherwise (Data File 5; Johnson et al., 2017).

MAAB Pipeline

The MAAB pipeline (Fig. 3) was executed within a KNIME workflow (Berthold et al., 2008). Metrics from most stages (for retained sequences) are included in MAAB output files (Data File 1 [Phytozome] and Data Files 3 and 4 [1KP]). The MAAB pipeline consists of two stages: stage 1, identification of HRGPs and removal of chimeric HRGPs; and stage 2, primary classification based on amino acid bias, motif analysis, and final classification (Fig. 3). Stage 1 consists of six steps, 1a to 1f. Stage 1a is a clustering step and removes sequences with 95% or greater identity; 1b calculates the percentage of each amino acid and the totaled amino acid biases, percentage PAST, percentage PSKY, and percentage PVKY (retained if one or more are 45% or greater); 1c removes all sequences of fewer than 90 amino acid residues (redundant step for 1KP data; see below); 1d retains all sequences of 10% or greater Pro; 1e identifies Conserved Domain Database (CDD) domains using NCBI RPSBLAST+, using the CDD database as of April 2013, E value cutoff of 1e-5, NCBI BLAST+ version 2.2.29, and retains those without a CDD domain. The CDD database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) includes some but not all models from PFAM (http://pfam.xfam.org/) and does not include EXT models, thereby ensuring their retention for further analysis. Stage 1f uses SignalP version 4.0 with default settings, with the exception of selecting the No-TM option and retaining only those sequences with an N-terminal ER signal sequence. For large data sets (1KP), a prefiltering step was added for increased efficiency to select sequences by amino acid bias, percentage Pro, and length (90 or more amino acids).

Stage 2 starts with primary classification based on amino acid bias and places sequences into one of four categories: AGP, if percentage PAST ≥ percentage PSKY+2 and percentage PVKY+2; EXT, if percentage PSKY ≥ percentage PAST +2 and percentage PVKY+2; PRP, if percentage PVKY ≥ percentage PAST+2 and percentage PSKY+2; and sequences with no clear amino acid bias (Δ amino acid bias < 2) are put into the shared bias category. Validation and final classification uses motif counting. The motifs used for AGPs are [ASVTG]P, [ASVTG]PP, and [AVTG]PPP; those used for EXT are SP_3, SP_4, SP_5, [FY]XY, KHY, VY[HKDE], VxY, and YY; and those used for PRPs are PPV[QK], PPVx[KT], and KKPCPP. Motif percentages also are calculated to facilitate ease of comparison (Fig. 4) but are not used for classification. The motifs were developed by identifying motifs from a broad class of published AGPs, EXTs, and PRPs, including those from sequenced protein backbones (Supplemental Table S1). A search for CL-EXT motifs (SP_n and then Y-based motifs) is done first, followed by PRP motifs (PPVx[KT], PPV[QK], and KKPCPP), and, finally, AGP motifs. Motifs are only accepted if there is no overlap with a motif accepted earlier during the motif search. Next, a total motif count is done, and all sequences with 15% or greater known motifs are taken through to a relative motif count. The relative HRGP motif count ensures that sequences have the motifs expected for the amino acid bias class they are categorized into. Validation of the primary classification is designated YES if the number of accepted motifs is greater than or equal to the accepted motifs for the other two classes (for AGPs and PRPs only). The number of accepted AGP motifs is calculated from the number of AGP motifs divided by 2, to account for the shortness of the AGP motif ([ASVTG]P). Accepted CL-EXT motifs have a minimum requirement of two SP_{3-5} motifs and two Y motifs. An additional requirement for CL-EXT classification is that SP_n and Y motifs must be present in a similar ratio (SP_n:Y between 0.25 and 4). These criteria ensure that CL-EXTs, with SP_n glycomotifs and interspersed Y-based cross-linking motifs, can be distinguished from volvocine algal HRGPs, which tend to have large domains containing only SP_n motifs. Some inconsistencies occur in the minor MAAB classes (6–8, 10–13, and 15–18) due in part to the order-dependent nature of the assignment of classes and implementation of the SP_n:Y ratio, especially when SP_n and/or Y counts are 0 and/or 1. The shared bias class and the sequences that do not meet these criteria (NO) are analyzed separately in the last step of stage 2. Before this last stage, all sequences are analyzed for the presence of a C-terminal GPI-anchor using the big-PI plant predictor (Eisenhaber et al., 2003).

At the completion of stage 2, all sequences are categorized into one of 24 classes (see "Results and Discussion"; Table I). Classes 1, 2, and 3 represent the major known classes of HRGPs based on Arabidopsis. Thirteen proteins (0.015%) are classified as errors, due to a tie between criteria for motif bias (to five decimal places), and they were ignored. This could be resolved in the future by an examination of motif count and coverage (percentage) of protein sequence by motifs from each family. The identification of Pro-rich regions of interest (or coverage) was performed as described at http://services.plantcell.unimelb.edu.au/hrgp/roi.java and is included in the MAAB outputs (Data Files 1, 3, and 4). AGP, CL-EXT, and PRP motifs were identified (Stage 2: Motif analysis) using a Java program, which searches for motifs in a family-directed manner (in the order CL-EXT, PRP, and AGP motifs, then longest first); overlapping motifs are not permitted during counting (http://services.plantcell.unimelb.edu.au/hrgp/maab_stage2.java). Names reported for each MAAB class in outputs and summary files (e.g. Data Files 1, 3, 4, 5, and 6) are old names and differ from the correct final descriptive names used in Figures 3 and 4, Table I, and throughout this article. The changes are summarized here as correct name (names used in Data Files): class 2, CL-EXT (2 Classical EXT); classes 19 to 23, shared bias (HRGP bias); class 24, less than 15% known HRGP motifs (24 HRGP).

Venn Diagrams

Venn diagrams were produced, one for each HRGP class 1 to 4 (Supplemental Fig. S4), using combined data from 1KP MAAB output (Data Files 3 and 4) and generated using R/Bioconductor (Gentleman et al., 2004).

Simple Search Method for Finding Motifs in MAAB Class 24 Sequences

Class 24 sequences were obtained from MAAB output files (Data Files 3 and 4) by sorting HRGP class data in Microsoft Excel and copying to a new tab. Each motif was searched using the find all option and the result was divided by 2, because each sequence is reported twice per row in the MAAB output files (Data Files 3 and 4). Wildcard * is used rather than X. Results are reported in Table II.

Motif Heat Maps

Sequences from class 1 (GPI-AGP), class 2 (CL-EXT), and class 24 (less than 15% known motifs) were analyzed for glycomotifs XP_1, XP_2, XP_3, XP_4, and XP_5 (separately for each motif within each HRGP class, where X = all 20 amino acids), and the percentage of each glycomotif per sequence was determined as described at http://services.plantcell.unimelb.edu.au/hrgp/xp_motif.java. The data for XP_1, XP_3, XP_4, and X = A, T, S, Y, K, V, G, F, H, and L are summarized as heat maps in Supplemental Figure S5 and were generated using R/Bioconductor (Gentleman et al., 2004).

Data Access and Data Sets

All data files for this article (Data Files 1–5) and the companion article (Data Files 6–9; Johnson et al., 2017) are listed here and numbered consecutively to avoid confusion.

Data File 1

MAAB output from analysis of 15 Phytozome genomes. Data for each of the 24 HRGP classes are in a separate sheet (tab) in the .xls file. Metrics from most MAAB stages (for retained sequences) are included. For most sequences, the species identifier is in the sequence name as follows: Aquca (*Aquilegia coerulea*), AT (Arabidopsis), Bradi (*Brachypodium distachyon*), Eucgr (*Eucalyptus grandis*), Medtr (*Medicago truncatula*), Pp1 (*Physcomitrella patens*), Potri (*Populus trichocarpa*), LOC_Os (rice [*Oryza sativa*]), Si (*Setaria italica*), Sb (*Sorghum bicolor*), Glyma (*Glycine max*), Solyc (*Solanum lycopersicum*), and Vocar (*Volvox carteri*). *Selaginella moellendorffii* sequences have only a number format (e.g. 440674|PACid:15406499), and *Chlamydomonas reinhardtii* sequences are in two formats, either Cre01gxxxxx or gxxxxx.t1|PAC. Filename: phytozome_hrgp_20150514.xls.

Data File 2

MAAB input 1KP proteins identified from oases $k = 39/49/59/69$ assembled transcriptomes that are at least 90 amino acids in length, compositionally biased (percentage PAST/percentage PSKY/percentage PVYK ≥ 45), and at least 10% Pro. Filename: RD001_oases_k39thru69_proteins_20150612.csv.gz.

Data File 3

MAAB output of 1KP data for 1KP data sets from eudicots to monilophytes, inclusive. Metrics from most stages (for retained sequences) are included. Filename: SA001_MAAB-hits-May2014-higher-clades.xls.

Data File 4

MAAB output of 1KP data for 1KP data sets from lycophytes to algae, inclusive. Metrics from most MAAB stages (for retained sequences) are included. Filename: SA002_MAAB-hits-May2014-lower-clades.xls.

Data File 5

A list of all sequences eliminated from MAAB because they were from data sets that contain some contamination. Filename: CR002_hits-excluded.xls.

Data File 6

Mean number of HRGPs in each HRGP class (columns) for each 1KP data set (rows), calculated from MAAB output files (Data Files 3 and 4). Filename: SA003_MAAB-hits-summary-by-class-and-sample.xls.

Data File 7

DNA sequences for 1KP GPI-AGPs (class 1) detected by MAAB. Locus identifiers (Data Files 3 and 4) of class 1 GPI-AGPs were used to extract the DNA sequence from the appropriate multiple-*k*-mer assembly. Where more than one locus identifier is reported for a single protein (e.g. with different *k*-mers), only one DNA sequence is reported. Filename: 1kp_agp_incl_dnas_9-12-2015.xls.

Data File 8

DNA sequences for 1KP CL-EXTs (class 2) detected by MAAB. Locus identifiers (Data Files 3 and 4) of class 2 CL-EXTs were used to extract the DNA sequence from the appropriate multiple-*k*-mer assembly. Where more than one locus identifier is reported for a single protein (e.g. with different *k*-mers), only one DNA sequence is reported. Filename: class2_incl_dnas_9-12-2015.xls.

Data File 9

Sequences identified by HMMER model 1 (putative AtAGP6/11 orthologs). Filename: agp6_model1_hmm_hits_20150922.xls.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Identity of GPI-AGPs and CL-EXTs.

Supplemental Figure S2. Alignment of Arabidopsis CL-EXTs.

Supplemental Figure S3. Representation of sequences identified by MAAB in Phytozome with features highlighted.

Supplemental Figure S4. Venn diagram of the number of HRGPs reported in the four major HRGP classes using four *k*-mer sizes.

Supplemental Figure S5. Conservation of XP_1, XP_3, and XP_4 glycomotifs in MAAB output by 1KP group.

Supplemental Figure S6. Sequences used for phylogenetic analysis (in fasta format).

Supplemental Table S1. Experimental evidence for hydroxylation and glycosylation of native Hyp-rich glycoproteins.

Supplemental Table S2. Comparison of MAAB and BIO OHIO/manual curation classification of Arabidopsis AGPs, EXTs, and PRPs.

Supplemental Table S3. Comparison of MAAB and BIO OHIO 2.0/manual curation classification of *P. trichocarpa* AGPs, EXTs, and PRPs.

ACKNOWLEDGMENTS

We thank Gane Ka-Shu Wong and colleagues Douglas E. Soltis, Nicholas W. Miles, Michael Melkonian, Barbara Melkonian, Michael K. Deyholos, James Leebens-Mack, Carl J. Rothfels, Dennis W. Stevenson, Sean W. Graham, Xumin Wang, Shuangxiu Wu, J. Chris Pires, Patrick P. Edger, and Eric J. Carpenter, who selected species for analysis and collected plant and algal samples, extracted RNA for sequencing, organized transcripts into gene families, screened RNA sequencing data for contamination, generated the sequence data, and set up and maintained 1KP Web resources used to communicate data; Mark Chase from the Kew Royal Botanic Gardens, Markus Ruhsam and Philip Thomas at the Royal Botanic Garden Edinburgh, and Steven Joya from the University of British Columbia for supplying plant samples to 1KP; and Drs. Birgit and Frank Eisenhaber (Bioinformatics Institute A*Star Singapore) for providing us with a stand-alone version of big-PI plant predictor.

Received March 6, 2017; accepted April 21, 2017; published April 26, 2017.

LITERATURE CITED

Babu MM (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochem Soc Trans **44:** 1185–1200

Bennick A (1987) Structural and genetic aspects of proline-rich proteins. J Dent Res **66:** 457–461

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: The Konstanz Information Miner. *In* C Preisach, H Burkhardt, L SchmidtThieme, R Decker, eds, Data Analysis, Machine Learning and Applications. Springer, Berlin, pp 319–326

Buljan M, Frankish A, Bateman A (2010) Quantifying the mechanisms of domain gain in animal proteins. Genome Biol **11:** R74

Chaturvedi P, Singh AP, Batra SK (2008) Structure, evolution, and biology of the MUC4 mucin. FASEB J **22:** 966–981

Coimbra S, Costa M, Jones B, Mendes MA, Pereira LG (2009) Pollen grain development is compromised in Arabidopsis *agp6 agp11* null mutants. J Exp Bot **60:** 3133–3142

Coimbra S, Costa M, Mendes MA, Pereira AM, Pinto J, Pereira LG (2010) Early germination of *Arabidopsis* pollen in a double null mutant for the arabinogalactan protein genes *AGP6* and *AGP11*. Sex Plant Reprod **23:** 199–205

Datta K, Schmidt A, Marcus A (1989) Characterization of two soybean repetitive proline-rich proteins and a cognate cDNA from germinated axes. Plant Cell **1:** 945–952

Draeger C, Ndinyanka Fabrice T, Gineau E, Mouille G, Kuhn BM, Moller I, Abdou MT, Frey B, Pauly M, Bacic A, et al (2015) Arabidopsis leucine-rich repeat extensin (LRX) proteins modify cell wall composition and influence plant growth. BMC Plant Biol **15:** 155

Dunker AK, Bondos SE, Huang F, Oldfield CJ (2015) Intrinsically disordered proteins and multicellular organisms. Semin Cell Dev Biol **37:** 44–55

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:** 1792–1797

Eisenhaber B, Wildpaner M, Schultz CJ, Borner GHH, Dupree P, Eisenhaber F (2003) Glycosylphosphatidylinositol lipid anchoring of plant proteins: sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol **133:** 1691–1701

Ellis M, Egelund J, Schultz CJ, Bacic A (2010) Arabinogalactan-proteins: key regulators at the cell surface? Plant Physiol **153:** 403–419

Estévez JM, Kieliszewski MJ, Khitrov N, Somerville C (2006) Characterization of synthetic hydroxyproline-rich proteoglycans with arabinogalactan protein and extensin motifs in Arabidopsis. Plant Physiol **142:** 458–470

Ferris PJ, Waffenschmidt S, Umen JG, Lin H, Lee JH, Ishida K, Kubo T, Lau J, Goodenough UW (2005) Plus and minus sexual agglutinins from *Chlamydomonas reinhardtii*. Plant Cell **17:** 597–615

Fincher GB, Stone BA, Clarke AE (1983) Arabinogalactan-proteins: structure, biosynthesis and function. Annu Rev Plant Physiol **34:** 47–70

Fong C, Kieliszewski MJ, de Zacks R, Leykam JF, Lamport DT (1992) A gymnosperm extensin contains the serine-tetrahydroxyproline motif. Plant Physiol **99:** 548–552

Forman-Kay JD, Mittag T (2013) From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure **21:** 1492–1499

Fowler TJ, Bernhardt C, Tierney ML (1999) Characterization and expression of four proline-rich cell wall protein genes in Arabidopsis encoding two distinct subsets of multiple domain proteins. Plant Physiol **121:** 1081–1092

Franssen HJ, Nap JP, Gloudemans T, Stiekema W, Van Dam H, Govers F, Louwerse J, Van Kammen A, Bisseling T (1987) Characterization of cDNA for nodulin-75 of soybean: a gene product involved in early stages of root nodule development. Proc Natl Acad Sci USA **84:** 4495–4499

Gaspar YM, Nam J, Schultz CJ, Lee LY, Gilson PR, Gelvin SB, Bacic A (2004) Characterization of the Arabidopsis lysine-rich arabinogalactan-protein *AtAGP17* mutant (rat1) that results in a decreased efficiency of *Agrobacterium* transformation. Plant Physiol **135:** 2162–2171

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Godl K, Hallmann A, Wenzl S, Sumper M (1997) Differential targeting of closely related ECM glycoproteins: the pherophorin family from *Volvox*. EMBO J **16:** 25–34

Gomord V, Fitchette AC, Menu-Bouaouiche L, Saint-Jore-Dupas C, Plasson C, Michaud D, Faye L (2010) Plant-specific glycosylation patterns in the context of therapeutic protein production. Plant Biotechnol J **8:** 564–587

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40:** D1178–D1186

Hall BG (2013) Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol **30:** 1229–1235

Held MA, Tan L, Kamyab A, Hare M, Shpak E, Kieliszewski MJ (2004) Di-isodityrosine is the intermolecular cross-link of isodityrosine-rich extensin analogs cross-linked in vitro. J Biol Chem **279:** 55474–55482

Hong JC, Nagao RT, Key JL (1987) Characterization and sequence analysis of a developmentally regulated putative cell wall protein gene isolated from soybean. J Biol Chem **262:** 8367–8376

Hong JC, Nagao RT, Key JL (1990) Characterization of a proline-rich cell wall protein gene family of soybean: a comparative analysis. J Biol Chem **265:** 2470–2475

Johnson KL, Cassin AM, Lonsdale A, Wong GK-S, Soltis DE, Miles NW, Melkonian M, Melkonian B, Deyholos MK, Leebens-Mack J, et al (2017) Insights into the evolution of hydroxyproline-rich glycoproteins from 1000 plant transcriptomes. Plant Physiol **174:** 904–921

Johnson KL, Jones BJ, Bacic A, Schultz CJ (2003a) The fasciclin-like arabinogalactan proteins of Arabidopsis: a multigene family of putative cell adhesion molecules. Plant Physiol **133:** 1911–1925

Johnson KL, Jones BJ, Schultz CJ, Bacic A (2003b) Non-enzymic cell wall (glyco)proteins. *In* J Rose, ed, The Plant Cell Wall. Blackwell Publishing, Oxford, pp 111–154

Johnson MTJ, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, Carrigan CT, Chase MW, Clarke ND, Covshoff S, Depamphilis CW, et al (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS ONE **7:** e50226

Khan IK, Kihara D (2016) Genome-scale prediction of moonlighting proteins using diverse protein association information. Bioinformatics **32:** 2281–2288

Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM (2015) Polymorphism analysis reveals reduced negative selection and elevated rate of insertions and deletions in intrinsically disordered protein regions. Genome Biol Evol **7:** 1815–1826

Kieliszewski M, de Zacks R, Leykam JF, Lamport DT (1992) A repetitive proline-rich protein from the gymnosperm Douglas fir is a hydroxyproline-rich glycoprotein. Plant Physiol **98:** 919–926

Kieliszewski MJ, Lamport DTA (1994) Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. Plant J **5:** 157–172

Kurotani A, Sakurai T (2015) In silico analysis of correlations between protein disorder and post-translational modifications in algae. Int J Mol Sci **16:** 19812–19835

Lamport DTA, Kieliszewski MJ, Chen Y, Cannon MC (2011) Role of the extensin superfamily in primary cell wall architecture. Plant Physiol **156:** 11–19

Lee KJD, Sakata Y, Mau SL, Pettolino F, Bacic A, Quatrano RS, Knight CD, Knox JP (2005) Arabinogalactan proteins are required for apical cell extension in the moss *Physcomitrella patens*. Plant Cell **17:** 3051–3065

Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-, and internal regions. Genome Inform Ser Workshop Genome Inform **10:** 30–40

Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A (2013) Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol **30:** 2645–2653

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK (2006) Intrinsic disorder in transcription factors. Biochemistry **45:** 6873–6888

Liu X, Wolfe R, Welch LR, Domozych DS, Popper ZA, Showalter AM (2016) Bioinformatic identification and analysis of extensins in the plant kingdom. PLoS ONE **11:** e0150177

Ma H, Zhao J (2010) Genome-wide identification, classification, and expression analysis of the arabinogalactan protein gene family in rice (*Oryza sativa* L.). J Exp Bot **61:** 2647–2668

Majewska-Sawka A, Nothnagel EA (2000) The multiple roles of arabinogalactan proteins in plant development. Plant Physiol **122:** 3–10

Manconi B, Castagnola M, Cabras T, Olianas A, Vitali A, Desiderio C, Sanna MT, Messana I (2016) The intriguing heterogeneity of human salivary proline-rich proteins. J Proteomics **134:** 47–56

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, et al (2014) Data access for the 1,000 Plants (1KP) project. Gigascience **3:** 17

Matsushima N, Tanaka T, Kretsinger RH (2009) Non-globular structures of tandem repeats in proteins. Protein Pept Lett **16:** 1297–1322

Matsushima N, Yoshida H, Kumaki Y, Kamiya M, Tanaka T, Izumi Y, Kretsinger RH (2008) Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins. Curr Protein Pept Sci **9:** 591–610

Motose H, Sugiyama M, Fukuda H (2004) A proteoglycan mediates inductive interaction during plant vascular development. Nature **429:** 873–878

Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet **14:** 157–167

Newman AM, Cooper JB (2011) Global analysis of proline-rich tandem repeat proteins reveals broad phylogenetic diversity in plant secretomes. PLoS ONE **6:** e23167

Nido GS, Méndez R, Pascual-García A, Abia D, Bastolla U (2012) Protein disorder in the centrosome correlates with complexity in cell types number. Mol Biosyst **8:** 353–367

Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztányi Z, Uversky VN, Obradovic Z, Kurgan L, et al (2013) D^2P^2: database of disordered protein predictions. Nucleic Acids Res **41:** D508–D516

Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins (Suppl 7) **61:** 176–182

Peng Z, Mizianty MJ, Kurgan L (2014) Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins **82:** 145–158

Pereira LG, Coimbra S, Oliveira H, Monteiro L, Sottomayor M (2006) Expression of arabinogalactan protein genes in pollen tubes of *Arabidopsis thaliana*. Planta **223:** 374–380

Pietrosemoli N, García-Martín JA, Solano R, Pazos F (2013) Genome-wide analysis of protein disorder in *Arabidopsis thaliana*: implications for plant environmental adaptation. PLoS ONE **8:** e55524

Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, et al (2017) DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res **45:** D1123–D1124

Radivojac P, Obradović Z, Brown CJ, Dunker AK (2003) Prediction of boundaries between intrinsically ordered and disordered protein regions. Pac Symp Biocomput 216–227

Rana SB, Zadlock FJ IV, Zhang Z, Murphy WR, Bentivegna CS (2016) Comparison of de novo transcriptome assemblers and k-mer strategies using the killifish, *Fundulus heteroclitus*. PLoS ONE **11:** e0153104

Rashid A (2014) Sub-cellular localization of PELPK1 in *Arabidopsis thaliana* as determined by translational fusion with green fluorescent protein reporter. [In Russian] Mol Biol (Mosk) **48:** 300–305

Rashid A, Deyholos MK (2011) PELPK1 (At5g09530) contains a unique pentapeptide repeat and is a positive regulator of germination in *Arabidopsis thaliana*. Plant Cell Rep **30:** 1735–1745

Roberts K (1974) Crystalline glycoprotein cell walls of algae: their structure, composition and assembly. Philos Trans R Soc Lond B Biol Sci **268:** 129–146

Roberts S, Dzuricky M, Chilkoti A (2015) Elastin-like polypeptides as models of intrinsically disordered proteins. FEBS Lett **589:** 2477–2486

Romero O, Obradovic, Dunker K (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Inform Ser Workshop Genome Inform **8:** 110–124

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins **42:** 38–48

Saha P, Ray T, Tang Y, Dutta I, Evangelous NR, Kieliszewski MJ, Chen Y, Cannon MC (2013) Self-rescue of an EXTENSIN mutant reveals alternative gene expression programs and candidate proteins for new cell wall assembly in *Arabidopsis*. Plant J **75:** 104–116

Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B (2011) Protein disorder: a breakthrough invention of evolution? Curr Opin Struct Biol **21:** 412–418

Schultz CJ, Ferguson KL, Lahnstein J, Bacic A (2004) Post-translational modifications of arabinogalactan-peptides of *Arabidopsis thaliana*: endoplasmic reticulum and glycosylphosphatidylinositol-anchor signal cleavage sites and hydroxylation of proline. J Biol Chem **279:** 45503–45511

Schultz CJ, Harrison MJ (2008) Novel plant and fungal AGP-like proteins in the *Medicago truncatula-Glomus intraradices* arbuscular mycorrhizal symbiosis. Mycorrhiza **18:** 403–412

Schultz CJ, Hauser K, Lind JL, Atkinson AH, Pu ZY, Anderson MA, Clarke AE (1997) Molecular characterisation of a cDNA sequence encoding the backbone of a style-specific 120 kDa glycoprotein which has features of both extensins and arabinogalactan proteins. Plant Mol Biol **35:** 833–845

Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A (2002) Using genomic resources to guide research directions: the arabinogalactan protein gene family as a test case. Plant Physiol **129:** 1448–1463

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28:** 1086–1092

Seifert GJ, Roberts K (2007) The biology of arabinogalactan proteins. Annu Rev Plant Biol **58:** 137–161

Shimizu M, Igasaki T, Yamada M, Yuasa K, Hasegawa J, Kato T, Tsukagoshi H, Nakamura K, Fukuda H, Matsuoka K (2005) Experimental determination of proline hydroxylation and hydroxyproline arabinogalactosylation motifs in secretory proteins. Plant J **42:** 877–889

Showalter AM, Basu D (2016a) Extensin and arabinogalactan-protein biosynthesis: glycosyltransferases, research challenges, and biosensors. Front Plant Sci **7:** 814

Showalter AM, Basu D (2016b) Glycosylation of arabinogalactan-proteins essential for development in Arabidopsis. Commun Integr Biol **9:** e1177687

Showalter AM, Keppler B, Lichtenberg J, Gu D, Welch LR (2010) A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. Plant Physiol **153:** 485–513

Showalter AM, Keppler BD, Liu X, Lichtenberg J, Welch LR (2016) Bioinformatic identification and analysis of hydroxyproline-rich glycoproteins in Populus trichocarpa. BMC Plant Biol **16:** 229

Shpak E, Barbar E, Leykam JF, Kieliszewski MJ (2001) Contiguous hydroxyproline residues direct hydroxyproline arabinosylation in *Nicotiana tabacum*. J Biol Chem **276:** 11272–11278

Starrett J, Garb JE, Kuelbs A, Azubuike UO, Hayashi CY (2012) Early events in the evolution of spider silk genes. PLoS ONE **7:** e38084

Sun X, Rikkerink EH, Jones WT, Uversky VN (2013) Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology. Plant Cell **25:** 38–55

Szalkowski AM, Anisimova M (2011) Markov models of amino acid substitution to study proteins with intrinsically disordered regions. PLoS ONE **6:** e20488

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol **30:** 2725–2729

Tan L, Leykam JF, Kieliszewski MJ (2003) Glycosylation motifs that direct arabinogalactan addition to arabinogalactan-proteins. Plant Physiol **132:** 1362–1369

Tseng IC, Hong CY, Yu SM, Ho THD (2013) Abscisic acid- and stress-induced highly proline-rich glycoproteins regulate root growth in rice. Plant Physiol **163:** 118–134

Varadi M, Guharoy M, Zsolyomi F, Tompa P (2015) DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder. BMC Bioinformatics **16:** 153

Velasquez SM, Marzol E, Borassi C, Pol-Fachin L, Ricardi MM, Mangano S, Juarez SP, Salter JD, Dorosz JG, Marcus SE, et al (2015) Low sugar is not always good: impact of specific *O*-glycan defects on tip growth in Arabidopsis. Plant Physiol **168:** 808–813

Voigt J, Frank R, Wöstemeyer J (2009) The chaotrope-soluble glycoprotein GP1 is a constituent of the insoluble glycoprotein framework of the Chlamydomonas cell wall. FEMS Microbiol Lett **291:** 209–215

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol **337:** 635–645

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci USA **111:** E4859–E4868

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, et al (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-seq reads. Bioinformatics **30:** 1660–1666

Yang J, Sardar HS, McGovern KR, Zhang Y, Showalter AM (2007) A lysine-rich arabinogalactan protein in Arabidopsis is essential for plant growth and development, including cell division and expansion. Plant J **49:** 629–640

Yang J, Zhang Y, Liang Y, Showalter AM (2011) Expression analyses of AtAGP17 and AtAGP19, two lysine-rich arabinogalactan proteins, in *Arabidopsis*. Plant Biol (Stuttg) **13:** 431–438

Yang Y, Smith SA (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics **14:** 328