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 disordered 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.

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 (Bułjan 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 PIOvos 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 (PIetroseomi 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; PIetroseomi 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; Majewksa-Sawka and Nothnagel, 2000; Seiffert 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 SO4,5 repeats that occur in EXTs (Fig. 1). In the highly glycosylated AGPs, large type II arabinino-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
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 SP₃ motifs), and a combination of both methods for PRPs. Many EXTs also include Y-based CL motifs, and searching for SP₃₅ 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 subclasses 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.

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) SP₁₅ 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 SP₁₅ 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); (2) 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.
(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.onkp.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 volvocene 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). AtAGPs 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 prollyl-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: AP1-3 (yellow); SP1-2 (blue); SPPP (also found in EXTs; blue underlined); TP1-3 (pink/purple); [G/V]P1-3 (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 included for comparison. 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 (SP3, SP4, SP5, or a combination), short, and chimeric (LRXs, PERKs, and others). Of the 59 EXTs reported by Showalter et al. (2010), we redefined SP3, SP4, and SP5 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 pairwise similarity is generally higher than for the GPI-AGPs (compare Supplemental Fig. S1, A and B).

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

<table>
<thead>
<tr>
<th>Species/Clade</th>
<th>BLASTp</th>
<th>MAAB Class</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species/Clade</td>
<td>GPI-AGP</td>
<td>Non-GPI-AGP</td>
<td>AGP</td>
</tr>
<tr>
<td>Species/Clade</td>
<td>CL-EXT</td>
<td>GPI-EXT</td>
<td>EXT</td>
</tr>
<tr>
<td>Species/Clade</td>
<td>PRP</td>
<td>GPR Bias</td>
<td>EXT Bias</td>
</tr>
<tr>
<td>Species/Clade</td>
<td>&lt;15% Motif</td>
<td>Shared Bias</td>
<td>Total</td>
</tr>
<tr>
<td>Species/Clade</td>
<td>Class 1–4</td>
<td>Total Class 5–23</td>
<td>Total Class 1–24</td>
</tr>
<tr>
<td>Monocots/commelinids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>1</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>4</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>2</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>3</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>Core eudicots/asterids</td>
<td>8</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>7</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>16</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Eucalyptus grandis</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Glycine max</td>
<td>7</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>12</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Basal eudicots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquilegia coerulea</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Lycophytes</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Selaginella moellendorfii</td>
<td>5</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Mosses</td>
<td>1</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Physcomitrella patens</td>
<td>1</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Green algae</td>
<td>2</td>
<td>78</td>
<td>1</td>
</tr>
<tr>
<td>Volvox carteri</td>
<td>1</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>2</td>
<td>15</td>
<td>71</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>14</td>
<td>66</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>3</td>
<td>48</td>
</tr>
</tbody>
</table>


eArabidopsis 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. bThe 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.
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

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 SP3, SP4, SP5, [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 SP₃₋₅ motifs and two Y motifs that must be present in a similar ratio (SP₃₋₅: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.

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
Figure 4. (Figure continues on following page.)
combinations of HRGP motifs. For example, in class 5 (AGP bias, high EXT [SPn and Y]), EXT30 (At1G02405) has few AG-glycomotifs and more EXT motifs with at least two SPn 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 SPn 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 over-representation of AGP motifs (class 10), SPn motifs (class 11, SPn:Tyr \( \geq \) 4), Y motifs (class 12, SPn:Tyr < 0.25), or PRP motifs (class 13). Sequences such as EXT39 (At5G19800) with EXT bias but only one SPn and/or one Y motif cannot be typical CL-EXTs with predominantly alternating SPn and Y motifs and, therefore, were included in class 20 (shared bias, high EXT [SPn 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 \( \Delta \)) 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 SPn/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 SPn/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. Another possibility is that more PRP motifs need to be incorporated into the MAAB pipeline. It is likely that such motifs have been

---

**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. SPn:Y is reported as the number of SPn motifs:number of Y motifs (ratio of SPn:Y reported as a fraction). White text for the SPn:Y ratio indicates that the sequence does not satisfy at least one of the criteria for CL-EXT: at least two SPn, and two Y motifs (indicated by asterisks) or a ratio of SPn: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 SP3-5 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.
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 Phytozone Proteomes

Additional analyses performed on the Arabidopsis data set downloaded from Phytozone 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 FAST (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 SRp 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 (At5G9530.1, 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 Phytozone 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).
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 assembled 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).

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

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 streptophyte algae (Johnson et al., 2017).
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.

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

<table>
<thead>
<tr>
<th>Motifs identified</th>
<th>Polymer</th>
<th>Vascular Plants&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Non-vascular Plants&lt;sup&gt;b&lt;/sup&gt;</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEPK</td>
<td>OsRePRP1</td>
<td>165</td>
<td>22</td>
<td>Tseng et al. (2013)</td>
</tr>
<tr>
<td>PEPKPKPEPK</td>
<td>OsRePRP1</td>
<td>17</td>
<td>0</td>
<td>Tseng et al. (2013)</td>
</tr>
<tr>
<td>PELPK</td>
<td>AtPRP10/AtPELPK1</td>
<td>17 4</td>
<td>Rashid and Deyholos (2011)</td>
<td></td>
</tr>
<tr>
<td>PEPK</td>
<td>AtPRP10/AtPELPK1</td>
<td>1,563 776</td>
<td>Rashid and Deyholos (2011)</td>
<td></td>
</tr>
<tr>
<td>(PEPK)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>AtPRP10/AtPELPK1</td>
<td>346 4</td>
<td>Rashid and Deyholos (2011)</td>
<td></td>
</tr>
<tr>
<td>KPPP</td>
<td>120-kD Douglas fir extensin&lt;sup&gt;c&lt;/sup&gt;</td>
<td>855 309</td>
<td>Fong et al. (1992); Schultz et al. (1997)</td>
<td></td>
</tr>
<tr>
<td>PGQGQQ</td>
<td>Gluten, wheat</td>
<td>0 1</td>
<td>Roberts et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>PPPVHL</td>
<td>γ-Zein</td>
<td>1</td>
<td>1</td>
<td>Matsumisha et al. (2008)</td>
</tr>
<tr>
<td>(PPG)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Collagen</td>
<td>1 8</td>
<td>Matsumisha et al. (2008)</td>
<td></td>
</tr>
<tr>
<td>(VPGXA)&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Elastin, human</td>
<td>266 639</td>
<td>Roberts et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>(VPGXA)&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Elastin, human</td>
<td>2 2</td>
<td>Roberts et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>PGMG</td>
<td>Biomaterialization molecule, sea urchin</td>
<td>8 16</td>
<td>Matsumisha et al. (2008)</td>
<td></td>
</tr>
</tbody>
</table>

### 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...
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 LÂTE 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; Hall, 2013). Phylogenetic analysis was performed at http://www.pondr.com using three different predictors: VI-XT (red; Romero et al., 1997, 2001; Li et al., 1999), VSL2 (green; Obradovic et al., 2005), and VL3 (blue; Radijovic 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.
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. plantcollaborative.org/wiki/display/iptool/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 RBPSLAST+, 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 PAST ≥ 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 [ASVTGP], [ASVTG]P, and [ASVTG]PP; those used for EXT are SP, SP, SP, FP, KY, KH, VY[HKDE], VX, and YY; and those used for PRPs are PPV[QK]. PPV[KT], and KKCPP. 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, and then Y-based motifs) is done first, followed by PRP motifs (PPV[KT], PPV[QK], and KKCPP), 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 SP3 motifs and two Y motifs. An additional requirement for CL-EXT classification is that SP and Y motifs must be present in a similar ratio (SP/Y between 0.25 and 4). These criteria ensure that CL-EXTs, with SP glycomotifs and interspersed Y-based cross-linking motifs, can be distinguished from volvocine algal HRGPs, which tend to have large domains containing only SP 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/Y ratio, especially when SP, 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 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 * was 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 XP1, XP2, XP3, XP4, and XP5 (separately for each motif within each HRGP class, where X = 20 amino acids), and the percentage of each glycomotif per sequence was determined as described at http://services.plantcell.unimelb.edu.au/hrgp/sp_motif.java. The data for XP1, XP2, XP3, and XP5 = A, T, S, Y, K, V, G, F, H, and L are summarized as heat maps in Supplemental Figure 55 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 coerulci).
AT (Arabidopsis), Bradi (Brachypodium distachyon), Eucgr (Eucalyptus grandis), Medtr (Medicago truncatula), Pp1 (Physcomitrella patens), Potri (Populus trichocarpa), LOC_Os (rice [Oryza sativa]), Si (Selaginella italica), Sb (Sorghum bicolor), Glyma (Glycine max), Solyc (Solanium lycopersicum), and Vcar (Volvocaceae). Selaginella moellendorfii sequences have only a number format (e.g. 440674|PACid:15406499), and Chlamydomonas reinhardtii sequences are in two formats, either Cre01gxxxx or gxxxxx.t1|PAC. Filename: phytozone_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 $\geq 45$), 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 agg_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 Phytozone 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$_3$, XP$_5$, and XP$_7$ 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 Rubsam 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


