Insights into the Evolution of Hydroxyproline-Rich Glycoproteins from 1000 Plant Transcriptomes


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.); Departments of Biological Sciences and Medicine, University of Alberta, Edmonton, Alberta, Canada, and BGI-Shenzhen, Bei Shan Industrial Zone, Yantian District, Shenzhen, China (G.K.-S.W., J.C.P.); Florida Museum of Natural History, Department of Biology, University of Florida, Gainesville, Florida 32611 (D.E.S., N.W.M.); Botanical Institute, Cologne Biocenter, University of Cologne, D50674 Cologne, Germany (M.M., B.M.); Department of Biology, University of British Columbia, Kelowna, British Columbia V1V 1V7, Canada (M.K.D.) Department of Plant Biology, University of Georgia, Athens, Georgia 3062 (J.L.-M.); University Herbarium and Department of Integrative Biology, University of California, Berkeley, California 94720 (C.J.R.); New York Botanical Garden, Bronx, New York 10458 (D.W.S.); Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada (S.W.G.); Key Laboratory of Genome Science and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China (X.W., S.W.); Division of Biological Sciences and Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211 (J.C.P.); Department of Horticulture, Michigan State University, East Lansing, Michigan 48823 (P.P.E.); 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.); 0000-0001-6108-5560 (G.K.-S.W.); 0000-0003-4205-3891 (M.K.D.); 0000-0003-4811-2231 (J.L.-M.); 0000-0002-6605-1770 (C.J.R.); 0000-0002-2986-7076 (D.W.S.); 0000-0001-8209-5231 (S.W.G.); 0000-0003-0302-8909 (X.W.); 0000-0001-9682-2639 (J.C.P.); 0000-0001-7483-8605 (A.B.); 0000-0002-8921-2725 (M.S.D.); 0000-0003-2026-9122 (C.J.S.).

The carbohydrate-rich cell walls of land plants and algae have been the focus of much interest given the value of cell wall-based products to our current and future economies. Hydroxyproline-rich glycoproteins (HRGPs), a major group of wall glycoproteins, play important roles in plant growth and development, yet little is known about how they have evolved in parallel with the polysaccharide components of walls. We investigate the origins and evolution of the HRGP superfamily, which is commonly divided into three major multigene families: the arabinogalactan proteins (AGPs), extensins (EXTs), and proline-rich proteins. Using motif and amino acid bias, a newly developed bioinformatics pipeline, we identified HRGPs in sequences from the 1000 Plants transcriptome project (www.onepkm.com). Our analyses provide new insights into the evolution of HRGPs across major evolutionary milestones, including the transition to land and the early radiation of angiosperms. Significantly, data mining reveals the origin of glycosylphosphatidylinositol (GPI)-anchored AGPs in green algae and a 3- to 4-fold increase in GPI-AGPs in liverworts and mosses. The first detection of cross-linking (CL)-EXTs is observed in bryophytes, which suggests that CL-EXTs arose though the juxtaposition of preexisting SP1 EXT glycomotifs with refined Y-based motifs. We also detected the loss of CL-EXT in a few lineages, including the grass family (Poaceae), that have a cell wall composition distinct from other monocots and eudicots. A key challenge in HRGP research is tracking individual HRGPs throughout evolution. Using the 1000 Plants output, we were able to find putative orthologs of Arabidopsis pollen-specific GPI-AGPs in basal eudicots.

Cell walls of plants and algae are widely used for food, textiles, paper, and timber, yet our understanding of their assembly and dynamic remodeling in response to growth, development, and environmental stresses (abiotic and biotic) remains rudimentary (Doblin et al., 2010, 2014). There are two contrasting types of walls/extracellular matrices in the green plant lineage, protein-rich walls of some green algae and the cellulose-rich walls of embryophytes (land plants). Recent studies of cell wall evolution suggest that the origins of many wall components occurred in the streptophyte green algae prior to the evolution of embryophytes (Sørensen et al., 2010; Popper et al., 2011; Domozych et al., 2012) with elaboration of a preexisting set of polysaccharides rather than an entirely new polymer framework. Although both the ultrastructure of plant walls and the fine structure of their component polymers vary widely, they all typically constitute a fibrillar network of cellulose microfibrils that is embedded within a gel-like matrix of noncellulosic polysaccharides, pectins,
and (glyco)proteins, the latter being both structural and enzymatic (Soromville et al., 2004; Doblin et al., 2010).

Research to date has largely focused on the evolution of the polysaccharides, cellulose, and the noncellulosic polysaccharides (hemicelluloses) and pectins, primarily through the availability of polysaccharide epitope-specific antibodies coupled with immunofluorescence and/or transmission electron microscopy and arraying techniques such as comprehensive microarray polymer profiling and plate-based arrays (Moller et al., 2007; Pattathil et al., 2010; Moore et al., 2014). In contrast, relatively little is known about the origin and evolution of the hydroxyproline-rich glycoproteins (HRGPs), a major group of wall glycoproteins, despite their widespread occurrence (Supplemental Table S1) and importance in plant growth and development (for review, see Fincher et al., 1983; Kieliszewski and Lamport, 1994; Majewska-Sawka and Nothnagel, 2000; Ellis et al., 2010; Lamport et al., 2011; Draeger et al., 2015; Velasquez et al., 2015; Showalter and Basu, 2016). Examination of these glycoproteins in a wider range of plant species should provide valuable insights into how they have evolved in parallel with the polysaccharide components in plant walls.

HRGPs are commonly divided into three major multigene families, the highly glycosylated arabino-galactan proteins (AGPs), the moderately glycosylated extensins (EXTs), and the minimally glycosylated proline-rich proteins (PRPs). The protein backbones of these HRGPs are distinguished by differences in amino acid bias and the repeat motifs characteristic to each family and include the posttranslational modification of Pro to Hyp (Johnson et al., 2017). Each of these families is further elaborated by chimeric and hybrid HRGPs, complicating not only the classification of members but also the elucidation of their functions.

Classical AGPs consist of up to 99% carbohydrate due to the addition of large type II arabino-3,6-galactan polysaccharides (degree of polymerization, 30–120) onto the minor (1%–10%) protein backbone. The complexity of this family is immense and lies both in the diversity of protein backbones containing AGP glycomotifs and the incredible heterogeneity of glycan structures and sugars decorating these proteins. The specific binding of AGPs to the dye β-glucosyl Yariv reagent (β-Glc Yariv) and the generation of AGP-glycan-specific antibodies have provided valuable insight into their function, structure, and evolution (Supplemental Table S1; Jermyn and Yeow, 1975; van Holst and Clarke, 1985; Kitazawa et al., 2013; Paulsen et al., 2014). These studies suggest that AGPs are evolutionarily ancient. O-Glycosylation occurs in bryophytes, some chlorophytes, and brown algae, as shown by the binding of β-Glc Yariv and/or the isolation of Hyp-containing glycoproteins (Godl et al., 1997; Ligrone et al., 2002; Lee et al., 2005; Shibaya and Sugawara, 2009; Hervé et al., 2016; Johnson et al., 2017). Classical AGPs are likely to exist in the chlorophyte algae, as one of the first-cloned HRGP genes (accession no. AAB23258.2) from Chlamydomonas reinhardtii, C. reinhardtii (Woessner and Goodenough, 1989), is predicted to be a glycosylphosphatidylinositol (GPI)-AGP according to our criteria (Johnson et al., 2017). It is likely that AGPs in the chlorophyte algae have distinct functions from streptophytes, as no large arabino-3,6-galactan polysaccharides have been reported (Ferris et al., 2001). Short, branched oligosaccharides (Ferris et al., 2001) and linear, Ara- and Gal-containing oligosaccharides (Bollig et al., 2007) have been isolated from C. reinhardtii walls.

The highly glycosylated AGPs are soluble and have been proposed to act as mobile signals, potentially involved in cell-cell communication (Schultz et al., 1998; Motose et al., 2004; Pereira et al., 2014). Linking of AGPs to the plasma membrane by a GPI-anchor is significant, as it points toward their increased organization within the cell wall-plasma membrane interface. The finding of GPI-AGPs in lipid rafts/nanodomains in eudicots is intriguing, as these have been proposed to stabilize and increase the activity of proteins and protein complexes at the plasma membrane involved in cell-cell communication, signal transduction, immune responses, and transport (Borner et al., 2005; Grennan, 2007; Tapken and Murphy, 2015). It is tempting to speculate on a role for GPI-AGPs in similar processes in all species in which they occur, although experimental evidence is largely lacking.
AGPs have been most studied in embryophytes and have been isolated from a wide range of tissues, including seeds, roots, stems, leaves, and inflorescences, and are particularly abundant in the stems of gymnosperms and angiosperms (Gaspar et al., 2001). The co-occurrence of specific glycan epitopes with developmental stages, the expression pattern of AGP genes, and mutant studies have implicated AGPs in many processes. These include roles in plant growth and development, cell expansion and division, hormone signaling, embryogenesis of somatic cells, differentiation of xylem, and responses to abiotic stress. The heterogeneity of the AGP family and the vast number of family members suggest that they are multifunctional, similar to what is found in mammalian proteoglycan/glycoproteins and intrinsically disordered proteins (IDPs; Filmus et al., 2008; Schaefer and Schaefer, 2010; Tan et al., 2012). Given that diverse members could have the same glycosylation and, conversely, that the same backbone could have alternate glycosylated forms, determining the function of any single AGP and the establishment of precise AGP function is a major challenge. Despite the magnitude of this task, enormous progress has been made. For example, two well-characterized AGPs are the pollen-specific AtAGP6 and AtAGP11, which play important roles in pollen development and female tissue interactions and are suggested to be involved in multiple, integrated signaling pathways (Nguema-Ona et al., 2012). Investigation of when these AGPs evolved would aid us in ascertaining their specific functions in pollen development.

The EXTs are an equally diverse family of HRGPs whose members contain repetitive SO3-5 motifs that direct moderate glycosylation with Ara oligosaccharides (arabinosides; degree of polymerization, 1–4). Classical EXTs play important structural roles through cross-linking into the wall, facilitated by Tyr (Y)-based motifs such as VXY and XYX (Tierney and Varner, 1987; Lamport et al., 2011). This cross-linking into the wall can have essential roles in plant development, such as the requirement for Arabidopsis (Arabidopsis thaliana) AtEXT3 during embryogenesis for cell plate formation (Cannon et al., 2008) and AtEXT18 in pollen cell wall integrity (Choudhary et al., 2015). Cross-linking of EXTs into the wall is proposed to result in the cessation of cell expansion, and yet, some EXTs are important for cell elongation in root hairs (Velasquez et al., 2011, 2015). The importance of glycosylation on EXT-like motifs in the moss Physcomitrella patens was shown recently through mutants in Hyp-O-arabinosyltransferases (MacAlister et al., 2016). Reduced levels of cell wall-associated Hyp arabinosides (as found in EXTs) in P. patens resulted in increased elongation of protonemal tip cells. Therefore, different classes of EXTs have evolved different biological roles, and these functions are governed by both their protein backbone and their glycosylation sequences and patterns.

The first EXTs were isolated from tomato (Solanum lycopersicum; Lamport, 1977), and proteins with blocks of SO3-5 motifs have been extracted from other angiosperms such as tobacco (Nicotiana tabacum), carrot (Daucus carota), sugar beet (Beta vulgaris), and alfalfa (Medicago sativa; Chen and Varner, 1985; Smith et al., 1986; Li et al., 1990) as well as other green plant lineages, including gymnosperms (Fong et al., 1992), the likely sister group to angiosperms, and chlorophycean alga (Woessner and Goodenough, 1989), representing the sister to the remaining green plants. EXTs from these other green plant lineages can have distinct features from embryophyte EXTs, and we use the term cross-linking (CL)-EXTs to distinguish EXTs with alternating SO3 and Y-based motifs, which occur predominantly in land plants, from the HRGPs that have a Y residues in a different domain from the SO3-rich domains, as occurs in volvocine algae (see Fig. 1 in Johnson et al., 2017). Exceptions may exist; for example, no classical EXTs were detected in the Hyp-poor cell walls of maize (Zea mays), which also do not have detectable isodityrosine, an intramolecular cross-link that occurs between Y residues in EXTs (Kieliszewski et al., 1990; Kieliszewski and Lamport, 1994).

The PRPs are the most difficult family of HRGPs to define, as very few have been studied to date. Investigation of PRPs has largely been restricted to legumes and Arabidopsis, where they have been shown to be minimally glycosylated, if at all (Averyhart-Fullard et al., 1988; Datta et al., 1989; Kleis-San Francisco and Tierney, 1990; Lindstrom and Vodkin, 1991; Fowler et al., 1999). The majority of PRPs are chimeric, as they contain an Ole PFAM domain, and they have been implicated in roles in root development (Bernhardt and Tierney, 2000), stress responses (Li et al., 2016), fiber development (Xu et al., 2013), and nodule formation (Sherrier et al., 2005). Despite these important functions, information on the origins and evolution of PRPs remains fundamentally unexplored.

Collectively, the functional characterization of HRGPs points toward their involvement in many processes within the wall, including signaling and cell wall integrity pathways (Costa et al., 2013; Pereira et al., 2015, 2016; Voeux and Höfte, 2016). As plants evolved multicellularity and transitioned to land, it is likely that greater complexity in signaling networks was required, a role that could seemingly be fulfilled by HRGPs, due to their distinct protein and glycan characteristics. It is currently unknown if the occurrence of specific HRGP is associated with evolutionary milestones and the specialization of plant cell walls. Greater insight into the origins, evolution, and diversity of HRGPs will enable researchers to address these key issues.

To study HRGP function throughout plant evolution, we developed a robust method for their detection that exploits the amino acid bias of the predicted protein backbone and characteristic glycomotifs (Johnson et al., 2017). The motif and amino acid bias (MAAB) pipeline classifies HRGP sequences into one of 24 predefined descriptive subclasses (23 HRGP classes and one MAAB class) and is optimized for the recovery of GPI-AGPs, non-GPI-AGPs, and CL-EXTs. Here, we use this automated bioinformatics pipeline to identify and
classify HRGPs within the extensive 1000 Plant (1KP) transcriptome data.

The 1KP project (www.onekp.com) has generated large-scale transcriptomic data for over 1000 plant and algal species, thereby providing deeper sampling of key plant taxa beyond those with sequenced genomes (Johnson et al., 2012; Matasci et al., 2014; Wickett et al., 2014). Species selected include mostly non-model land plants, many of which are important for medicine, agriculture, forestry, and biodiversity/conservation, as well as red algae (Rhodophyta), green algae (chlorophytes and algal streptophytes), and taxa from a small group of freshwater unicellular algae known as glaucophytes, collectively termed the Archaeplastida (Plantae). Analyses of the 1KP data are already helping resolve long-standing questions, such as from which streptophyte green algal lineage the progenitors of land plants likely arose (Wickett et al., 2014). The 1KP transcriptome data set, therefore, offers an unprecedented opportunity to investigate HRGP superfamily evolution and determine if changes coincide with major evolutionary transitions in the Archaeplastida.

A tandem repeat annotation library (TRAL) is also used to identify repeats in MAAB sequences and support findings of selective loss and/or gain of CL-EXTs in bryophytes and commelinid monocots. A combination of BLAST and profile hidden Markov models was used to search for putative orthologs of pollen-specific AtAGP6/11 in angiosperms (flowering plants) with hits validated by phylogenetic analysis. These vignettes demonstrate the broad utility of the bioinformatics pipeline on the 1KP data, providing evolutionary insights into the complexity of the HRGP superfamily.

RESULTS AND DISCUSSION

MAAB Pipeline for GPI-AGPs and CL-EXTs

Using the newly developed bioinformatics tool, the MAAB pipeline (Johnson et al., 2017), we evaluated the 1KP data set for HRGPs. We set parameters in MAAB to best detect the classical, non-chimeric GPI-AGPs, CL-EXTs, PRPs, and non-GPI-AGPs (classes 1–4, respectively) as well as the hybrid HRGPs (classes 5–23) using features of the well-characterized Arabidopsis sequences. MAAB was validated using 15 proteomes from Phytozone and shown to reliably detect and accurately classify HRGPs into 23 descriptive subclasses (Johnson et al., 2017). MAAB represents a significant advance of previous methods to detect HRGPs in being able to identify all classes in one pipeline, without manual curation.

MAAB did not identify any false positives in Phytozone data sets, and evaluation of a subset of the 1KP data sets also suggests a very low rate of false positives (Johnson et al., 2017). In Ranunculaceae, a well-sampled basal eudicot order (39 samples derived from 20 species), we analyzed 143 non-redundant GPI-AGP sequences (five Ranunculaceae and six Papaveracea data sets; Supplemental Fig. S1A) and observed 98.6% true positives, with only two sequences not being clear GPI-AGPs (class 1) and four other sequences flagged as likely contaminants based on their appearance in multiple unrelated data sets (see legend, Supplemental Fig. S1). The rate of false positives for CL-EXTs (class 2) was estimated based on analysis of 45 non-redundant bryophyte sequences (Supplemental Fig. S2) derived from the larger set of 67 identified within the 1KP data (Fig. 1). Shading of motifs reveals that 42 of 45 sequences (93.3%) have the expected CL-EXT motifs throughout the entire backbone. The three remaining sequences could be considered CL-EXTs, or hybrid CL-EXTs-AGPs or AGPs, as they have more AGP motifs than SP motifs, but both motif types are present (Supplemental Fig. S2). In either case, they are definitive HRGPs, suggesting that the true positive rate is 100% (45/45), although we cannot rule out that the sequences are partial chimeric HRGPs until full-length sequences become available.

Using the MAAB Pipeline to Identify and Classify HRGPs in Transcriptomic Data Sets Spanning Large Evolutionary Time Scales

Across the entire 1KP data, 45,304 HRGPs (classes 1–23) were identified (Fig. 1) using a multiple k-mer approach (Johnson et al., 2017). The majority of sequences detected were in HRGP classes 1 to 4 (38,051) and MAAB class 24 (39,933), with a grand total of 85,237 sequences across all 24 MAAB classes. We compared HRGPs identified for each of the 1KP taxonomic groups (hereafter 1KP groups, which includes clades and grades; see http://www.onekp.com/samples/list.php) by calculating the percentage of sequences in the classical (classes 1–4) and minor (classes 5–23) HRGP classes and the final MAAB class (class 24) out of the total sequences classified by MAAB (Fig. 2A). The majority of 1KP groups had a similar proportion of sequences in classes 1 to 4 (average ~50%), 5 to 23 (average ~10%), and 24 (average ~40%). However, the algal groups, including those well represented in the 1KP data set (green algae, comprising chlorophytes and algal streptophytes, and red algae and Chromista), exhibited a trend toward lower percentages of HRGP sequences (classes 1–4 and 5–23) and a higher percentage of sequences in MAAB class 24 (Fig. 2A). The relatively low number of GPI-AGPs (class 1) and the higher number of class 4 sequences most likely reflect underlying functional differences in algal AGPs. The high number of class 24 sequences, likely either non-HRGPs or unknown HRGPs, suggests that algae have a more diverse array of IDP’s than land plants. The possibility that HRGP transcripts have not been detected within k-mer assemblies also is possible and could be due to sequence quality and/or sampling issues (lower than optimal number of data sets and tissues; see below).
As a measure of the depth of taxon sampling, the number of taxonomic orders represented and the number of samples within each 1KP group are reported in Figure 2B. The green alga group is the best represented of the algae, with 34 orders and 152 samples included in our analyses, and hence provides the most comprehensive data set for comparison. The average number of sequences for all MAAB classes (1–24) is summarized by the 1KP group (Fig. 2C). As each group represents a variable number of species, and not all species in a group have the same HRGPs, we have also summarized the detection rate as a heat map by the 1KP group (Fig. 2C) and by taxonomic order (Supplemental Table S2). This allowed us to assess whether detection of a particular HRGP class is either real or spurious. If 50% or more species in a given taxonomic order have one or more sequence for a given HRGP class, we report a greater than 50% detection rate and assume that this HRGP class is widespread throughout the taxonomic order. The few class 2 (CL-EXTs) sequences identified in green algae and Chromista are likely sporadic contaminants based on the low mean number of CL-EXTs (0.1 for green algae and 0.2 for Chromista; Fig. 2C), low detection rates (Fig. 2C; Supplemental Table S2), and BLASTp analysis revealing high sequence similarity to vascular plant CL-EXTs (data not shown).

Here, we present a preliminary analysis of the minor HRGP classes (5–23) and will then focus on the major HRGP classes (1–4). Classes 5 to 23 include hybrid-type HRGPs with motifs associated with two or more classes of HRGP (see Fig. 4 in Johnson et al., 2017). For example, class 5 HRGPs contain AGP bias with CL-EXT motifs (both SP3-5 and Y-based) and are most highly represented in gymnosperms (conifers, Gnetales, Ginkgo, and Cycadales; Fig. 2C). In general, the detection of classes 5 to 23 is low and sporadic for most clades, with the exception of conifers and core eudicots. In these clades, a high detection rate (greater than 50%) of sequences occurs for most HRGP classes (Fig. 2C). This suggests that more hybrid/unusual HRGPs occur in these clades, although the low mean numbers indicate that these sequences are found in a few, rather than all, data sets within core eudicot and conifer orders. Data sets from the 1KP green algae group have a relatively high number of sequences in class 6 (1.4 ± 0.4; AGP bias, EXT glycomotifs [SPn], but not Y-based cross-linking motifs; Fig. 2C), which is consistent with our current knowledge of HRGPs isolated from Chlamydomonas and Volvox (Adair et al., 1983; Woessner and Goodenough, 1989; Woessner et al., 1994; Ferris et al., 2005). The majority of CL-EXTs identified in Arabidopsis are not predicted to be GPI-anchored (see Table I and Supplemental Table S2 in Johnson et al., 2017), and consistent with this, only a few GPI-anchored EXTs (class 9) were observed in Phytozome and 1KP data. Only one class, GPI-anchored PRPs (class 14), had no data reported for HRGP class 14. Selected data are reported for assemblies with k-mer = 25 (k25; shaded columns). Red numbers indicate that the sequences detected within this class and 1KP group are likely to be contaminants (see “Results and Discussion”).

As a measure of the depth of taxon sampling, the number of taxonomic orders represented and the number of samples within each 1KP group are reported in Figure 2B. The green alga group is the best represented of the algae, with 34 orders and 152 samples included in our analyses, and hence provides the most comprehensive data set for comparison. The average number of sequences for all MAAB classes (1–24) is summarized by the 1KP group (Fig. 2C). As each group represents a variable number of species, and not all species in a group have the same HRGPs, we have also summarized the detection rate as a heat map by the 1KP group (Fig. 2C) and by taxonomic order (Supplemental Table S2). This allowed us to assess whether detection of a particular HRGP class is either real or spurious. If 50% or more species in a given taxonomic order have one or more sequence for a given HRGP class, we report a greater than 50% detection rate and assume that this HRGP class is widespread throughout the taxonomic order. The few class 2 (CL-EXTs) sequences identified in green algae and Chromista are likely sporadic contaminants based on the low mean number of CL-EXTs (0.1 for green algae and 0.2 for Chromista; Fig. 2C), low detection rates (Fig. 2C; Supplemental Table S2), and BLASTp analysis revealing high sequence similarity to vascular plant CL-EXTs (data not shown).
representatives from any 1KP group, which is consistent with our current, albeit limited, knowledge of PRPs (Averyhart-Fullard et al., 1988; Datta et al., 1989; Fowler et al., 1999).

The mean numbers of GPI-AGPs and CL-EXTs in the 1KP data were lower than expected, particularly in the core eudicots/rosids (approximately four and seven respectively; Fig. 2C), compared with the numbers in the majority of eudicot and commelinid genomic data sets (see Table I in Johnson et al., 2017). We suspected that this might be due to the restricted number of tissue types used for sequencing, so we investigated a subset of 1KP data sets in the Ranunculales (39 samples derived from 20 species). For each data set, we manually removed likely redundant sequences (see "Materials and Methods"), reducing the mean number of GPI-AGP sequences from 7.3 to 4.8 (Supplemental Fig. S1A) and CL-EXTs from 5.1 to 2.9 (Supplemental Fig. S1B). Most of the redundant sequences contained variably sized and positioned deletions (data not shown).

Figure 2. Summary of HRGP sequences identified using the MAAB pipeline. A, Percentage of total HRGP sequences (by 1KP group) that are found in the classical HRGP classes (1–4), hybrid HRGP classes (5–23), and non-HRGPs (class 24). B, Number of orders analyzed and number of samples per order for each 1KP group. The monocot group excludes the commelinid monocots. C, Overview of the mean number of HRGPs for each MAAB class in the 1KP data set (by 1KP group). Shading of boxes represents the detection rate, indicated as the percentage of orders with hits (Supplemental Table S2). A shaded box showing a detection of HRGPs with a value of zero indicates an average number of sequences between 0 and 0.1. Red numbers indicate that the sequences detected within this HRGP class and 1KP group are likely to be contaminants (see "Results and Discussion").
The majority of the 1KP angiosperm data sets are from leaf tissue, yet this tissue has the lowest average number of GPI-AGPs (four), with 4.4 in roots, five in flowers, and six in developing fruits (Supplemental Fig. S1A). Leaf tissue data sets also had a low number of CL-EXTs (2.6) compared with roots (3.8) and fruits (3.3; Supplemental Fig. S1B). When all unique sequences are tallied across the four *Papaver* spp., 12 GPI-AGPs (Supplemental Fig. S1C) and six CL-EXTs were identified (Supplemental Fig. S1D). Thus, the total numbers of GPI-AGPs and CL-EXTs observed in most 1KP data sets are underestimates due to limited tissue sampling, despite the likely sequence redundancy that was not routinely removed by MAAB pipeline processing.

Significantly, the greater species sampling within 1KP mostly corroborated our Phytolome observations and allow us to draw some major conclusions: (1) AGPs are widespread, both GPI-AGPs and non-GPI-AGPs being convincingly found in the green algae group as well as Chromista and Glaucophyta; (2) CL-EXTs occur in most groups of extant land plants, including bryophytes (nonvascular plants); and (3) non-chimeric PRPs (class 3) are uncommon and largely confined to eudicots (Fig. 2C).

**GPI-AGPs Evolved Early in the Green Plant Lineage**

Our data show that GPI-AGPs are present in all algal groups sampled in the 1KP project with the exception of the Rhodophyta. The absence of GPI-AGPs in red algae is consistent with several other independent studies. A detailed analysis of the completed genome of two unicellular red algae, *Cyanidioschyzon merolae* (Hashimoto et al., 2009; Ulvskov et al., 2013) and *Galdiera sulphuraria* (Ulvskov et al., 2013), revealed that they lack the key enzymes for the synthesis of GPI-anchors; therefore, GPI-AGPs would not be expected to be present in these algae. To our knowledge, there are no reports in the literature of red algal AGPs based on either AGP glycan epitopes or β-Glc Yariv staining (Supplemental Table S1). AGPs are not well known in the Chromista, which include brown algae, although there are reports of low levels of Hyp (less than 0.04%) in the soluble fraction of cell walls from a variety of brown algal species (Gotelli and Cleland, 1968). Also, in a recent study, O-linked glycosylation typical of AGPs was detected in five species of brown algae, and analysis of the *Ectocarpus* proteome shows that proteins containing AGP-like glycomotifs are present (Hervé et al., 2016). A similar distribution of the amino acid motifs AP, SP, and TP is seen in all 1KP groups from Chromista to eudicots (see Supplemental Fig. S5 in Johnson et al., 2017), suggesting that AGP-type glycomotifs are evolutionarily ancient.

A consistent trend in genomic and transcriptomic data sets was the decrease in the number of non-GPI-AGPs from the green algae group to vascular plants and the opposite trend (an increase) for GPI-AGPs, including a 3- to 4-fold increase in GPI-AGPs, in liverworts and mosses (approximately four sequences per species), compared with the green algae group and hornworts (Figs. 1 and 2). The green algae (chlorophytes and algal streptophytes) are well represented in the 1KP data, including several species from Volvocaceae (two species, three samples) and Chlamydomonaceae (five species, five samples). The average number of GPI-AGPs and non-GPI-AGPs in the 1KP sample of *Chlamydomonas* spp. is 0.8 and 43.6, respectively, whereas in *Volvox* spp., it is one and 78.3, respectively (Data File 6). It is possible that some of the non-GPI-AGPs could be partial chimeric HRGPs rather than true non-GPI-AGPs; further work is needed to resolve this issue.

The functional specification of selected AGPs may have been established early in green plant evolution. For example, AGPs have been shown to be involved in apical growth in the moss *P. patens* (Lee et al., 2005) and some vascular plants (Nguema-Ona et al., 2012). β-Glc Yariv staining in eight liverwort species revealed ubiquitous tissue staining, with darker staining in apices of some species, suggesting that apical targeting of AGPs may be a feature that arose in bryophytes (hornworts, mosses, and liverworts; Basile and Basile, 1987). AGPs also have been implicated in cell plate formation in liverworts (Shibaya and Sugawara, 2009). Since apical targeting as well as AGP delivery to phragmoplasts (a scaffold for cell plate assembly) exist in streptophyte algae (Charales, Coleochaetales, and Zygmenatales), it is possible that these two functions of AGPs evolved earlier than the emergence of bryophytes (Leliaert et al., 2012; Bowman, 2013).

It should be noted, however, that β-Glc Yariv staining and glycan epitope labeling can provide only general information about the presence/absence of AGP-specific glycans, as these tools are not able to distinguish the different subclasses of AGPs. The information revealed by MAAB in the 1KP data detailing the suite of AGPs present in bryophytes in addition to the increasing use of the bryophyte models *P. patens* (Kofuji and Hasebe, 2014) and *Marchantia polymorpha* (Ishizaki et al., 2016) will facilitate molecular approaches to investigate the functions of specific AGPs through, for example, mutant studies.

**Origin and Selective Loss of CL-EXTs**

Previous studies of HRGPs show that EXT-like Hyparabinosides are evolutionarily ancient and are found in chlorophyte algae, such as *Chlorella vulgaris* (Lampert and Miller, 1971) and *C. reinhardtii* (Ferris et al., 2001; Bollig et al., 2007). Our data suggest that CL-EXTs arose in bryophytes, as no CL-EXTs were detected in the two volvocine (chlorophyte green algal) genomes (see Table I in Johnson et al., 2017) and only contaminating sequences were detected in algal 1KP data sets (Fig. 2). EXT epitopes have been identified in some chlorophyte algal species (Domozych et al., 2009; Sørensen et al., 2011; Supplemental Table S1), but this likely reflects the
presence of chimeric and/or hybrid EXTs rather than CL-EXTs. Similar to the Phytozome analysis, hybrid HRGPs with both AGP and EXT glycomotifs were found in the 1KP green algae group (e.g. RYJX_Locus_857 [Pandorina morum] and VALZ_Locus_7817 [C. noctigama], both class 19). Where Y residues are present, these are found in a separate domain from the P-rich domain (see Fig. 1 in Johnson et al., 2017). This result suggests that CL-EXTs arose though the juxtaposition of preexisting SPn EXT glycomotifs with refined Y-based motifs (e.g. YXY and VXY) rather than via the simultaneous evolution of both protein motifs.

To gain a better understanding of the occurrence of CL-EXTs in bryophytes, the phylogenetic distribution of CL-EXTs present in the 1KP data was assessed in more detail (Fig. 3). CL-EXTs were detected by the MAAB pipeline in all five hornwort species representing two taxonomic orders (Supplemental Fig. S2), whereas they were identified in only five of 16 orders of mosses (Hypnales, Pottiaceae, Diphysciaceae, Buxbaumiales, and Sphagnales) and two of eight orders of liverworts (Marchantiales and Sphaerocarpales) that were sampled by 1KP (Supplemental Table S2). This relatively low detection rate in mosses and liverworts is unlikely to be due to low species representation, as it is comparable to the sampling depth in hornworts (Fig. 3). The data, therefore, suggest that while CL-EXTs are widespread in hornworts, they have a more limited distribution within the moss and liverwort lineages. If hornworts are the sister group of other land plants (Wickett et al., 2014), then this may imply that CL-EXTs were lost from some lineages of liverworts and mosses. An alternative explanation is that CL-EXT genes are present in mosses and liverworts but are expressed at very low levels and, therefore, were not detected.

To exclude the possibility that MAAB missed partial CL-EXT sequences due to a strict criterion for an endoplasmic reticulum (ER) signal sequence (Johnson et al., 2017), a TRAL was generated to identify CL-EXT repeat motifs (Schaper et al., 2015). The TRAL identified all repeats in class 2 sequences that were longer than 10 amino acids, present in four or more copies, and significantly different from random sequence evolution ($P < 0.05$; see “Materials and Methods”). A library for MAAB class 24 sequences also was generated as a control. Each repeat was used to construct a circular sequence profile hidden Markov model, and these models were used to search the MAAB input data (see Data File 2 in Johnson et al., 2017) to identify all sequences that contain these repeats. A simple tool for browsing the results of TRAL is available at http://services.plantcell.unimelb.edu.au/hrgp/index.html, and this tool was used to search all the bryophyte species for TRAL hits.

There is a good correlation between the number of species with CL-EXTs from MAAB and those with SPn/Y-based repeats identified by TRAL (Fig. 3; Data File 6). Most importantly, there was no increase in the number of bryophyte species containing TRAL repeats (in the MAAB input data; see Data File 2 in Johnson et al., 2017) compared with MAAB class 2 sequences, suggesting that only a small number of partial CL-EXTs are excluded by the MAAB pipeline.

TRAL repeats from class 24 also were searched, as these could identify both novel HRGP motifs and other repeats of interest that may have been excluded by the MAAB-imposed filters. The majority of bryophyte species had hits to TRAL repeats generated from class 24 sequences; however, none contained both SPn and Y repeats and only a few have repeats containing Y, suggesting that class 24 contains few, if any, CL-EXTs. Combined, TRAL and MAAB outputs indicate that there are a number of moss and liverwort species that lack CL-EXTs. This leads us to suggest that there has been either selective loss (e.g. Jungermanniales and most species from Hypnales) or multiple independent gains (e.g. Sphagnales and Marchantiales) of CL-EXTs in mosses and liverworts. Confirming this result using genome sequencing of selected species and investigation of the morphology and development of mosses and liverworts with and without CL-EXTs would be an interesting avenue for future research. These non-vascular plants also may help uncover the role of CL-EXTs with different sequence motifs and repeat lengths.

**CL-EXTs Have Been Lost in Many Grass Species**

An evaluation of Phytozome data shows that CL-EXTs containing both SPn and Y-based motifs are found in almost all eudicot species but are surprisingly absent from Eucalyptus grandis (eudicot/rosid), Aquilegia coerulea (basal eudicot), and the four species of grasses (monocot/commelinid; Poaceae). No CL-EXTs were found in either P. patens or the two volvocine algal species (see Table I in Johnson et al., 2017). The low number of HRGPs in some of these species is suspected to be due to annotation and assembly issues, based on our detection of CL-EXT-like open reading frames in these genomes using 1KP-identified CL-EXTs as query sequences (see “Materials and Methods”). In contrast, this approach was not successful for genomes from the commelinid monocots (APG IV, 2016), designated as monocots/commelinids in the 1KP group nomenclature. This absence is convincing because there are 17 different grass species (21 data sets) in the 1KP data, and one species, Eleusine coracana, is represented by four different data sets (leaves, flowers, and unstressed and stressed roots; see Data File 3 in Johnson et al., 2017). As an additional check, we employed BUSCO (Benchmarking Universal Single-Copy Orthologs), an analysis tool that quantitatively assesses transcriptome completeness based on the presence of near-universal single-copy orthologous genes selected from OrthoDB (Simão et al., 2015). Between 65% and 92% (excluding three outliers) of 956 single-copy genes were identified in the monocots/commelinids data sets (Fig. 4B). The value for the Poaceae was at the high end of this range (median, 88.4%; mean, 83.8%), suggesting that the grass
transcriptome assemblies are high quality and that CL-EXTs should have been detected if they were present. Our findings also are supported by a recent bioinformatics study of the maize and rice (Oryza sativa) genomes, where no CL-EXTs were identified (Liu et al., 2016).

Primary cell walls of commelinid monocots have a composition distinct from eudicots such as Arabidopsis and gymnosperms (Bacic et al., 1988; Carpita and Gibeaut, 1993; Doblin et al., 2010). Therefore, we investigated whether an absence of CL-EXTs also was observed in other commelinid families represented by 1KP. As for Poaceae, CL-EXTs were not detected in the sedge family Cyperaceae (total of three species) and three other families within Poales, although they were identified in Bromeliaceae and Restionaceae (one species each; Fig. 4A). CL-EXTs also were not detected in two families within the Zingiberales (Zingiberaceae and Marantaceae). The commelinid monocots are a major clade within the larger monocot group (APG IV, 2016). As CL-EXTs are absent from at least four grass genomes, including the Sanger-sequence rice genome (see Table I in Johnson et al., 2017), and occur widely in the 1KP monocot group (which excludes the commelinid monocots; Fig. 2; Supplemental Table S2; Data File 6), we suggest that there has been selective loss of CL-EXTs within the commelinid monocots. Further studies are required to confirm their absence from specific commelinid species.

It is likely that the cross-linking function of CL-EXTs would need to be replaced by something else in commelinid monocots to provide the necessary scaffold for cell wall development. Obvious candidates are monomeric phenolic acids such as the hydroxycinnamic acids (ferulic and p-coumaric) of commelinid monocot walls, absent in the noncommelinid monocot, eudicot (with the exception of the Caryophyllales), and gymnosperm walls, which form covalent cross-links between arabinofuranosyl chains (for review, see Bacic et al., 1988; Harris, 2005). Additional experimentation is needed to test this hypothesis.

Orhologs of GPI-AGPs and CL-EXTs in Brassicales

The broad sampling of species in the 1KP data provides an opportunity to investigate how far orthologous relationships of selected sequences can be traced in plant lineages (Wickett et al., 2014). Given that IDPs...
are usually poorly conserved throughout evolution. We wanted to test if putative HRGP orthologs could be detected. As Arabidopsis HRGP sequences are the most well characterized, we undertook phylogenetic analysis of Arabidopsis and GPI-AGPs and CL-EXTs identified by MAAB in the Brassicales (see “Materials and Methods”; Fig. 5). The ML tree of the GPI-AGPs shows that sequences from Brassicaceae data sets group most closely with the Arabidopsis sequences, generally with good (greater than 70%) bootstrap values. Sequences from other Brassicales families cluster further away according to their evolutionary relatedness and with lower bootstrap support (Fig. 5A). Although showing low bootstrap support, AtAGP59, the additional GPI-AGP identified by MAAB analysis, clusters with AGP6 and AGP11 (Fig. 5A).

The phylogenetic analysis of Brassicales CL-EXTs revealed a similar hierarchical pattern to the GPI-AGPs, with Brassicaceae sequences generally clustering nearer to the Arabidopsis sequences, but with variable bootstrap support (Fig. 5B). Some CL-EXT 1KP sequences did not group with an Arabidopsis sequence, making the prediction of orthologous relationships difficult, even within the Brassicaceae. This is perhaps not surprising, since tandem repeat proteins often evolve more rapidly than other proteins (Moesa et al., 2012; van der Lee et al., 2014). The clustering pattern of GPI-AGPs in the Brassicales tree (Fig. 5A) largely reflects that observed for the Arabidopsis sequences (see Fig. 2C in Johnson et al., 2017). This suggests that, despite the low bootstrap values, it may be possible to identify orthologous sequences using phylogenetic analysis over larger evolutionary distances, such as within angiosperms (flowering plants).

Detection of Orthologs in Angiosperms: AtAGP6/11 as an Example

One pair of Arabidopsis GPI-AGPs, AtAGP6 and AtAGP11 (see subclade AGP-h in Fig. 2C in Johnson et al., 2017), has been well studied, as they are involved in pollen development and pollen tube growth and, therefore, impact reproductive capacity (Levitin et al., 2008; Coimbra et al., 2009, 2010). Therefore, orthologs of these genes could be expected to be found in a broader subset of eudicots and potentially early angiosperms, depending on gene conservation. Although a recent attempt to find putative orthologs of AtAGP6/11 in Quercus suber (cork oak; core eudicots/rosids, Fagales), pollen EST data were unsuccessful (Costa et al., 2015), we predicted that it should be possible to identify orthologs in the 1KP data, particularly within data sets that include floral tissue. A combination of BLAST (Phytozome and National Center for Biotechnology Information [NCBI]) and profile hidden Markov (HMMER) models (see “Materials and Methods”) was used to search for putative AtAGP6/11 orthologs in the 1KP multiple k-mer data (MAAB input and output; see Data Files 2–4 in Johnson et al., 2017). In some cases, additional sequences were identified by BLAST from other sources (e.g. the Amborella trichopoda genome).
Figure 5. Maximum likelihood (ML) tree of Brassicales 1KP and Arabidopsis GPI-AGPs (A) and CL-EXTs (B). ML trees generally show strong support for subclades with Arabidopsis sequences and 1KP sequences from family Brassicaceae. Putative GPI-AGP subclades (A) are separated by a horizontal dotted line, and the Arabidopsis orthologs are indicated by boxes (right of tree). Sequences from the Brassicaceae are in green, with Arabidopsis sequences in larger, boldface font.
for use as full-length evolutionary markers (see “Materials and Methods”). The HMMER models were optimized by varying the input sequences and the length of the aligned region (see “Materials and Methods”). The best results were obtained using full-length AtAGP6/11-type proteins (including the N-terminal ER and C-terminal GPI anchor signal sequences) from the Brassicaceae family (model 1; Supplemental Table S3). This was measured by a higher true positive and a lower false positive rate, as assessed by manual curation (Supplemental Table S4; see “Materials and Methods”).

The resulting sequences were subjected to phylogenetic analysis, and the ML tree displayed in Figure 6 shows the relatedness of a large subset of AtAGP6/11-like sequences derived from 1KP data and sequenced genomes. Almost all the eudicot sequences group together with AtAGP6/11. The monocot and commelinid monocot sequences form a separate subclade with AtAGP58 and AtAGP59 (Fig. 6), albeit with low bootstrap support. Notably, no subclade with AtAGP5 and AtAGP59 (Fig. 6), containing those containing floral tissue (Supplemental Table S4), despite two similar sequences being identified in the Quercus robur genome (Qurob_scaffold_362 and Qurob_scaffold_354). Only two sequences identified by HMMER model 1 were apparent false positives (OHAE_12357 and OHAE_4810; Polygala lutea, Fabales), positioned in a subclade with Arabidopsis AtAGP1/2/4/5 and OsAGP1 from rice (Fig. 6). The relatively clear separation of putative AtAGP6/11 orthologs from the other GPI-AGPs suggests that phylogenetic analysis may provide a reasonable prediction of gene orthology for these IDPs. In this case, potential AtAGP6/11 orthologs could be confidently detected in both eudicot and monocot lineages, suggesting the existence of an ancestral gene in a common ancestor prior to their divergence ~140 to 150 million years ago (Chaw et al., 2004). However, no potential AtAGP6/11 ortholog was identified in A. trichopoda, which is the sister group of all other angiosperms. Unfortunately, all eight 1KP data sets from basal angiosperms (members of Amborellaceae, Nymphaeales, and Austrobaileyales) were derived from either leaf or shoot tissue, not floral tissue, and A. trichopoda is the only basal angiosperm genome that is publicly available. Our demonstrated success at finding more HRGPs using multiple k-mer assemblies (Johnson et al., 2017) suggests that using this approach to analyze existing and future basal-most angiosperm floral transcriptome data sets would be advantageous before concluding an absence of AtAGP6/11 orthologs among this group.

To explore the robustness of the AtAGP6/11 subclade further, all GPI-AGPs from the 1KP data sets reported in Figure 6 were used to produce a larger angiosperm GPI-AGP tree (Supplemental Fig. S3). As expected, the bootstrap values were low; however, most of the additional 290 GPI-AGP sequences grouped into the subclades observed in Figure 2C of our companion article (Johnson et al., 2017). These reflect putative Arabidopsis orthologs for AtAGP17/18, AtAGP9, AtAGP4/7/10, AtAGP58, AtAGP25/26/27, AtAGP1, and AtAGP6/11/59. Importantly, despite the low bootstrap support, the majority of 1KP sequences in the AtAGP6/11 subclade in Figure 6 also group with AtAGP6/11 in the angiosperm GPI-AGP tree (Supplemental Fig. S3), suggesting that these sequences are indeed the most similar and, therefore, likely to be AtAGP6/11 orthologs within the analyzed transcriptomes.

Investigation of the GPI-AGP sequence alignment and further analysis of each individual GPI-AGP subclade revealed differences in the presence and distribution of specific amino acids (e.g. K, M, Q, and D/E) among subclade members (Fig. 7; Supplemental Fig. S4). For example, putative orthologs of AtAGP6/11 from both eudicots and some monocots share scattered K residues, concentrated in the first half of the protein followed by a region containing several acidic residues (Fig. 7; Supplemental Fig. S4K).

The scattered K residues are noticeably absent from the putative orthologs from Q. robur (Supplemental Fig. S4, A and K). Therefore, further analysis is required to confirm these sequences and other putative AtAGP6/11 genes in order to determine the true AtAGP6/11 orthologs. With additional taxonomic and appropriate tissue sampling, it should be possible to fill the evolutionary gaps and trace the origin of AtAGP6/11 genes as well as other HRGPs. Complementation of angiosperm mutant lines, such as the agp6 agp11 double mutant of Arabidopsis that displays pollen phenotypes (Levitin et al., 2008; Coimbra et al., 2009, 2010), with progenitor AtAGP6/11 genes will be essential to test their functional orthology and may reveal information about these motifs and, possibly, the specific glycosylation required for HRGP function in particular tissue types.
HRGPs: Remaining Challenges and Future Perspectives

Despite being minor components of the wall, the HRGPs make a significant contribution to wall properties and to cellular identity (Knox et al., 1991). With access to the 1KP data, we have made significant progress toward understanding the evolutionary time frame of when this large gene family of diverse and complex proteins likely evolved. We have provided a platform to guide experimental approaches to confirm our data and answer questions regarding how HRGPs evolved as well as to investigate the losses of major classes of HRGPs through evolution (e.g. loss of CL-EXTs in mosses, liverworts, and commelinid monocots).

Tracking individual HRGPs throughout evolution is a major challenge. Our investigation of AtAGP6/11 orthologs shows that it is possible to trace HRGPs with specific characteristics and determine their evolutionary origins. However, the CL-EXTs in particular, as a consequence of their variable tandem repeat motifs and diverse sequence length, cannot be aligned in a meaningful way with the tools developed for folded proteins. Alternative approaches are needed, such as investigating length variation and insertions/deletions rather than pairwise alignment. An approach that could be useful is to perform global multiple sequence alignments using a graph-based method that takes into account the observation that insertions/deletions can occur anywhere in a repeat (Szalckowski and Anisimova, 2013).

The biggest challenge will be to understand the individual and coordinated roles of HRGPs within the context of either a single cell or tissue type. Our data provide a valuable resource to initiate system-wide analyses to identify spatiotemporal expression changes of HRGPs throughout development. Even in the less complex nonvascular plants, GPI-AGPs and CL-EXTs are multigene families, and understanding sequences from other sources to increase the breadth of sampling (see “Materials and Methods”; Supplemental Table S4). The tree also includes at least one sequence from each Arabidopsis GPI-AGP subclade, representative rice GPI-AGPs identified by MAAB (see Fig. 2 in Johnson et al., 2017), and four GPI-AGPs from A. trichopoda (Amtri_ERM96654, Amtri_ERM95342, Amtri_ERN01113, and Amtri_ERN06202). Sequence names are colored by 1KP group: basal angiosperms (gray), non-commelinid monocots (purple), commelinid monocots (pink), basal eudicots (aqua), core eudicots (green), asterids (red), and rosids (orange). Numbers on the nodes represent support with 100 bootstrap replicates (70 or greater, green; 60–69 orange; 40–59, black). 1KP sequences are identified by Group_Order_1KP identifier_sequence locus. Genomic sequences and other sequences from NCBI follow a similar format but include a five-letter abbreviation of genus and species. Symbols next to sequence names are used to indicate the source and MAAB class of sequences and other relevant information. An asterisk after the sequence name indicates that additional information is provided in Supplemental Table S4; for example, the YNXR (Poales) data set is contaminated with Asparagales, although the ML tree suggests that these sequences are indeed Poales. The scale bar for branch length measures the number of substitutions per site.

Figure 6. Phylogenetic analysis of class I GPI-AGPs to identify AtAGP6/11 orthologs. The ML tree (MEGA) was constructed using putative AtAGP6/11 orthologs identified predominantly from 1KP transcriptomes and HmMER model 1 (Supplemental Table S3), with a few
their individual functions will provide exciting challenges for researchers for many years to come. Defining the role of individual AGPs and EXTs is confounded by their interactions with other cell wall polymers; cross-linking of AGPs (Tan et al., 2013) and EXTs to other polymers such as pectins and heteroxylans has been experimentally verified (for review, see Velasquez et al., 2012). Understanding how widespread cross-linking of HRGPs to other cell wall polymers is will be particularly challenging. The emerging availability of tractable genetic systems for non-model land plants will allow researchers to use the information provided by MAAB to explore these questions in a range of species.

Glycosylation is a characteristic feature of HRGPs and defines the interactive molecular surface. Predicting the posttranslational modification of HRGPs, including the sites of P hydroxylation and types of glycosylation, 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). For many HRGPs, glycosylation often completely encases and, hence, masks the protein backbone, as is the case for GPI-AGPs, arabino-3,6-galactan peptides, and CL-EXTs (see Fig. 1 in Johnson et al., 2017). Furthermore, it also drives the three-dimensional conformation of the protein backbone and, therefore, determines the presentation of epitopes on the molecule surface. Given that plant HRGPs have the same structural complexity as animal mucins, it is expected that the core protein sequence will influence the extent and type of glycosylation, which also can vary in a tissue-/cell-specific manner (Gerken et al., 2013). It is known that HRGP glycosylation differs between volvocine algae and embryophytes, but information on streptophyte green algae is lacking. Introducing the same HRGP backbone into multiple species and tissue types throughout evolution would be an interesting approach to investigate context-dependent glycosylation patterns and/or functions. A complementary investigation of the glycosyltransferases that act to glycosylate HRGPs also would greatly assist our understanding of this complex process. An exciting future of HRGP research beckons to determine the biological and functional significance of this fascinating superfamily of secreted glycoproteins.

**Figure 7.** Schematic representation of the GPI-AGP subclades showing the presence and distribution of specific amino acids in the PAST-rich protein backbone. The order of GPI-AGPs is based on clades in the angiosperm GPI-AGP tree (Supplemental Fig. S3); figure parts A to K are based on alignments (Supplemental Fig. S4), and the AtAGPS/10 (G) subclade is included for comparison (based on the AGP-c subclade; see Fig. 2C in Johnson et al., 2017). A, Putative orthologs of Lys-rich AtAGP17/18 have scattered Lys (K) residues throughout the sequence and contain a short K-rich domain near the C terminus. The glycomotifs are predominantly XP1-2. B, The AGP9 subclade also has a short Lys-rich region with three or more K residues and scattered K residues; however, the glycomotifs are distinct from AtAGP17/18, being predominantly [S/T]P3, C, D, and G to I, AGP4/7/10 (C), AGP2/3 (D), AGP5/10 (G), AGP5 (H), and AGP1 (I) subclades have a classical PAST-rich backbone with no obvious bias toward other amino acids. C, Most of the putative orthologs of Lys-rich AtAGP17/18 are relatively Met (M) rich, with scattered M residues between AGP glycomotifs and a Q residue at the presumed N terminus. This Q residue at the presumed mature N terminus also is found in putative orthologs of Lys-rich AtAGP17/18 (A), AtAGP9 (B), AtAGP1/2/3/4/7/5/10 (C, D, and G–I), and AtAGPS/8 (E) but not AtAGP25/26/27 (F) or AGP6/11/59 (J/K). J/K, Putative orthologs of AtAGP6/11/59 share scattered K residues, concentrated in the first half of the protein and also a small cluster of acidic residues, either Asp (D) or Glu (E).
MATERIALS AND METHODS

Phylogenetic Analysis

Phylogenetic analyses were performed using MEGA 6.06 (Tamura et al., 2013). Sequences (untrimmed) were first aligned using MUSCLE (Edgar, 2004) using the default settings, except for Supplemental Figure S3, where gap = -2 (default gap = -2.9). The best model was identified [find best protein model (ML) option] and used for each alignment to generate a maximum likelihood tree (with 100 bootstrap replications) using all sites in the alignment. Sequences used for phylogenetic analysis (fasta format), and the best model, are given in Supplemental Figure S3.

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; Johnson et al., 2017). Five additional data sets were analyzed (July 2015) to improve the coverage of hornworts (UCRN-Megaceros_tosanus, FAJP-Parapatamotoceros_halli, RXRQ-Paeoceros_carolinianus-sporophyte, and WCBZ-Paeoceros_carolinianus-gametophyte) and to include the single sample from clade Ginkgoales (SGTW-Ginkgo biloba). These additional five data sets are included in the MAAB input and output files (see Data Files 2-4 in Johnson et al., 2017), and most of the analysis with these samples is presented, with the exception of Supplemental Table S2.

MAAB Summary Analysis

The sequences and associated annotations from MAAB classification, including the assembly in which they were identified, are provided in Data Files 3 and 4 in the companion article (Johnson et al., 2017). Data from individual samples (hereafter referred to as data sets) that were identified by the 1KP consortium as containing contaminants were removed from analyses (unless noted otherwise) and can be found in Data File 5 (Johnson et al., 2017). The number of sequences in each MAAB class for each individual 1KP sample is provided in Data File 6. The mean number of HRGPs by MAAB class for each 1KP group [Fig. 2C] or commelinid monocot family [Fig. 4, A and C] was calculated by averaging the appropriate data in Data File 6 (which reports the number of sequences per HRGP class [columns] for each 1KP data set [rows]). Total numbers of HRGPs by 1KP group (Fig. 1) also were generated from Data File 6. Graphs (Fig. 4, A and C) were generated using R/Bioconductor (Gentleman et al., 2004). DNA sequences of GPI-AGPs (class 1) and CL-EXTs (class 2) are provided in Data Files 7 and 8 respectively.

Reducing Redundancy in 1KP Data Sets

For phylogenetic analyses, redundant sequences were eliminated from data sets. The most probable sequence was identified based on existing knowledge, as follows: MAAB 1KP output (see Data Files 3 and 4 in Johnson et al., 2017) was sorted by 1KP locus. Sequences for each data set (1KP locus) were then sorted by HRGP class, and the sequences from the desired HRGP class were pasted into a new Microsoft Excel sheet. Sequences were sorted alphabetically by sequence (to group sequences with similar N termini) and converted to fasta format. Multiple sequence alignments were performed with MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) to identify redundant sequences. Where two or more sequences had large regions of identity, the longer sequence was retained unless there were clear artifacts at either the N or C terminus. Retained sequences were subjected to further rounds of alignment to ensure that only unique sequences remained. When redundancy was unclear, both sequences were retained in small data sets but only one for larger data sets.

Removing redundancies from the 1KP sequence data was necessary because, like other short-read sequence data, these offers a wealth of information, but redundancy can occur from multiple sources, including splice variants, sequencing errors, recent duplications, chimeric transcripts, and polyplody (Gruenheit et al., 2012; Yang and Smith, 2013; Xie et al., 2014). Sequence redundancies were obvious in our analysis (Supplemental Fig. S1, A and B), and two relatively common occurrences were read through of GC-rich hairpins in the cDNA during reverse transcription and poor-quality sequence at one end of the paired read. cDNA synthesis for all 1KP data sets was performed at 42°C (Chen Li, personal communication).

TRAL Construction

TRAL version 0.3.5 (Schaper et al., 2015) was used to identify and evaluate tandem repeats present in 1KP sequences from HRGP class 2 (CL-EXTs), class 3 (PRPs), and class 24 (less than 15% known motifs), as these MAAB classes are the most likely to include tandem repeats. Tandem repeats were identified using T-REKS (Jorda and Kajava, 2009) and HHrepID (Biegert and Söding, 2008) algorithms. Tandem repeats were accepted only if they passed all the following criteria: (1) tandem repeat unit length of at least 10 amino acids; (2) at least four tandem repeat unit copies; (3) P < 0.05 using the phylo_gap01 model; and (4) repeat identified by T-REKS and/or HHrepID. A total of 5,139 repeats were identified and accepted from MAAB sequences in classes 2, 3, and 24. Each repeat was used to construct a circular sequence profile hidden Markov model (Schaper et al., 2015) and then searched with HMMER version 3.1b1 hmmsearch (E value cutoff of 1e-5; http://hmmer.org/) to identity repeats in MAAB input data (Data File 2 in Johnson et al., 2017). The resulting 4,070 models had two or more hits above threshold to the input data. These hits were used to construct alignments (maximum of 300 sequences per alignment), randomly chosen after 95% per sample clustering using USEARCH (Edgar, 2010), and aligned with MUSCLE version 3.81 (Edgar, 2004) for each model. A tool for browsing and searching TRAL results is available at http://services.plantcell.unimelb.edu.au/hrgp/index.html and was used to determine the number of species with class 2 and class 24 TRAL repeats, using keyword searches by order and/or species (Fig. 3).

BUSCO Analysis

1KP transcriptomes were assessed for completeness using the BUSCO analysis tool (Simão et al., 2015). The BUSCO plant data set and software were downloaded from http://buscos.ezlab.org/.

Profile Hidden Markov Models for Finding Putative AtAGP6/11 Orthologs

Protein sequences were aligned with MUSCLE version 3.81 (Edgar, 2004) using default parameter settings. Sequence alignments were then used as input into HMMER version 3.1b1 hmmbuild (http://hmmer.org/) to construct a profile hidden Markov model. Subsequently, hmmsearch was used (with default thresholds) to search against MAAB 1KP input sequences (see Data File 2 in Johnson et al., 2017). Seven different models were evaluated (Supplemental Table S3), and the results of the best model (see Data File 4 in Johnson et al., 2017) are summarized in Supplemental Table S4. For each positive 1KP data set, a single best sequence was selected manually based on presence of N-terminal ER signal or full or partial C-terminal GPI signal sequences and then checked by iterative multiple sequence alignment and phylogenetic analysis. To distinguish between true class 4 non-GPI-AGPs and potential partial GPI-AGPs, proteins needed to have the following GPI anchor signal characteristics: potential s0, s-upper sites, and s-lower sites, and a basic residue and/or a hydrophobic domain (Eisenhaber et al., 2003), unless noted otherwise (e.g. ZHMB_16956 and XMVD_14518), as summarized in Supplemental Table S4. Additional AtAGP6/11-like sequences were obtained from other sources (outlined below) for use as full-length controls and to increase the sequence representation in various 1KP groups. TBLASTn was used to search Phytozone and/or the nucleotide collection (NR/NT at NCBI) using word size 2, no filtering, and restricting organisms to same species, genus, or family as appropriate, with either AAAGP6/11 or OsAGP7/10 as query sequence. If unsuccessful, a suitable 1KP sequence was selected (coding sequence only; Data File 7) and BLASTn was performed (word size 7 and no filtering). When positive hits were identified, these were used as query sequences to find additional sequences (maximum of two per species used). Quercus robur AtAGP6/11 putative orthologs were identified in TBLASTn searches of Q. robur assembly version 1 scaffolds available at https://urgi.versailles.inra.fr/blasto. For the basil angiosperm Amborella trichopoda (not present in Phytozone version 9), GPI-AGPs were identified among annotated proteins accessible at ftp://ftpensemblgenomes.org/pub/plants/release-31/fasta/amborella_trichopoda/ using 50% PAST as outlined by Schultz et al. (2002). A similar approach was used to find evidence for genomic sequences of CL-EXTs from selected genomes using 1KP-identified CL-EXTs as query sequences (Data File 8). For example, 1KP sequence AYMT_Locus_34756 identified partial CL-EXT sequence Eucalyptus grandis scaffold_6: 36791844…36792375, and VCGH_Locus_4377 identified Aquilagia coerulea scaffold_34: 1845734…1846286.)
Data Access and Data Sets

Additional data sets for this article are listed below. Numbering continues from Johnson et al. (2017), and a full list of data files is provided there.

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. The locus identifier (Data Files 3 and 4) of class 1 GPI-AGPs was 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. The locus identifier (Data Files 3 and 4) of class 2 CL-EXTs was 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. Number of GPI-AGPs and CL-EXTs by tissue in Ranunculaceae and Papaveraceae samples represented in 1KP.

Supplemental Figure S2. Bryophyte CL-EXT sequences.

Supplemental Figure S3. Phylogenetic analysis (ML tree) of GPI-AGPs (all subclades) from selected angiosperms.

Supplemental Figure S4. Conserved features for putative orthologs of selected Arabidopsis GPI-AGPs.

Supplemental Figure S5. Sequences used for phylogenetic analysis (in fasta format).

Supplemental Table S1. Summary of selected AGP and EXT epitopes described in the literature.

Supplemental Table S2. HRGP detection rate for each 1KP taxonomic order.

Supplemental Table S3. Summary of HMMER models used to identify putative orthologs of AtAGP6s and AtAGP11.

Supplemental Table S4. Summary of HMMER model performance by 1KP samples and analysis of sequences selected for phylogenetic analysis.

ACKNOWLEDGMENTS

We thank 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 as well as Drs. Birgit and Frank Eisenhaber (Bioinformatics Institute A*Star Singapore) for providing us with a stand-alone version of the big-PI plant predictor.

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

LITERATURE CITED


Voxeur A, Höfte H (2016) Cell wall integrity signaling in plants: “To grow or not to grow that is the question.” Glycobiology 26: 950–960


