A Bioinformatics Approach to the Identification, Classification, and Analysis of Hydroxyproline-Rich Glycoproteins[W][OA]

Allan M. Showalter*, Brian Keppler, Jens Lichtenberg, Dazhang Gu, and Lonnie R. Welch

Molecular and Cellular Biology Program, Department of Environmental and Plant Biology (A.M.S., B.K.), and Center for Intelligent, Distributed, and Dependable Systems, Russ College of Engineering and Technology (J.L., D.G., L.R.W.), Ohio University, Athens, Ohio 45701–2979

Hydroxyproline-rich glycoproteins (HRGPs) are a superfamily of plant cell wall proteins that function in diverse aspects of plant growth and development. This superfamily consists of three members: hyperglycosylated arabinogalactan proteins (AGPs), moderately glycosylated extensins (EXTs), and lightly glycosylated proline-rich proteins (PRPs). Hybrid and chimeric versions of HRGP molecules also exist. In order to "mine" genomic databases for HRGPs and to facilitate and guide research in the field, the BIO OHIO software program was developed that identifies and classifies AGPs, EXTs, PRPs, hybrid HRGPs, and chimeric HRGPs from proteins predicted from DNA sequence data. This bioinformatics program is based on searching for biased amino acid compositions and for particular protein motifs associated with known HRGPs. HRGPs identified by the program are subsequently analyzed to elucidate the following: (1) repeating amino acid sequences, (2) signal peptide and glycosylphosphatidylinositol lipid anchor addition sequences, (3) similar HRGPs via Basic Local Alignment Search Tool, (4) expression patterns of their genes, (5) other HRGPs, glycosyl transferase, prolyl 4-hydroxylase, and peroxidase genes coexpressed with their genes, and (6) gene structure and whether genetic mutants exist in their genes. The program was used to identify and classify 166 HRGPs from Arabidopsis (*Arabidopsis thaliana*) as follows: 85 AGPs (including classical AGPs, lysine-rich AGPs, arabinogalactan peptides, fasciclin-like AGPs, plastocyanin AGPs, and other chimeric AGPs), 59 EXTs (including SP_5 EXTs, SP_5/SP_4 EXTs, SP_4 EXTs, SP_4/SP_3 EXTs, a SP_3 EXT, "short" EXTs, leucine-rich repeat-EXTs, proline-rich extensin-like receptor kinases, and other chimeric EXTs), 18 PRPs (including PRPs and chimeric PRPs), and AGP/EXT hybrid HRGPs.

The genomics era has produced vast amounts of biological data that await examination. In order to "mine" such data effectively, a bioinformatics approach can be utilized to identify genes of interest, subject them to various in silico analyses, and extract relevant biological information on them from various public databases. Examination of such data produces novel insights with respect to the genes in question and can be used to facilitate and guide further research in the field. Such is the case here, where bioinformatics tools were developed to identify, classify, and analyze members of the Hyp-rich glycoprotein (HRGP) superfamily encoded by the Arabidopsis (*Arabidopsis thaliana*) genome.

HRGPs are a superfamily of plant cell wall proteins that are subdivided into three families, arabinogalactan proteins (AGPs), extensins (EXTs), and Pro-rich proteins (PRPs), and extensively reviewed (Showalter, 1993; Kieliszewski and Lamport, 1994; Nothnagel, 1997; Cassab, 1998; José-Estanyol and Puigdomènech, 2000; Seifert and Roberts, 2007). However, it has become increasingly clear that the HRGP superfamily is perhaps better represented as a spectrum of molecules ranging from the highly glycosylated AGPs to the moderately glycosylated EXTs and finally to the lightly glycosylated PRPs. Moreover, hybrid HRGPs, composed of HRGP modules from different families, and chimeric HRGPs, composed of one or more HRGP modules within a non-HRGP protein, also can be considered part of the HRGP superfamily. Given that many HRGPs are composed of repetitive protein sequences, particularly the EXTs and PRPs, and many have low sequence similarity to one another, particularly the AGPs, BLAST searches typically identify only a few closely related family members and do not represent a particularly effective means to identify members of the HRGP superfamily in a comprehensive manner.

Building upon the work of Schultz et al. (2002) that focused on the AGP family, a new bioinformatics software program, BIO OHIO, developed at Ohio University, makes it possible to search all 28,952 proteins encoded by the Arabidopsis genome and identify putative HRGP genes. Two distinct types of searches are possible with this program. First, the program can search for biased amino acid compositions in the

* Corresponding author; e-mail showalte@ohio.edu.

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: Allan M. Showalter (showalte@ohio.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.110.156554

genome-encoded protein sequences. For example, classical AGPs can be identified by their biased amino acid compositions of greater then 50% Pro (P), Ala (A), Ser (S), and Thr (T), as indicated by greater than 50% PAST. Similarly, arabinogalactan peptides (AG peptides) are identified by biased amino acid compositions of greater then 35% PAST, but the protein (i.e. peptide) must also be between 50 and 90 amino acids in length. Likewise, PRPs can be identified by a biased amino acid composition of greater then 45% PVKCYT. Second, the program can search for specific amino acid motifs that are commonly found in known HRGPs. For example, SP_4 pentapeptide and SP_3 tetrapeptide motifs are associated with EXTs, a fasciclin H1 motif is found in fasciclin-like AGPs (FLAs), and PPVX(K/T) (where X is any amino acid) and KKPCPP motifs are found in several known PRPs (Fowler et al., 1999). In addition to searching for HRGPs, the program can analyze proteins identified by a search. For example, the program checks for potential signal peptide sequences and glycosylphosphatidylinositol (GPI) plasma member anchor addition sequences, both of which are associated with HRGPs (Showalter, 1993, 2001; Youl et al., 1998; Sherrier et al., 1999; Svetek et al., 1999). Moreover, the program can identify repeated amino acid sequences within the sequence and has the ability to search for bias amino acid compositions within a sliding window of user-defined size, making it possible to identify HRGP domains within a protein sequence.

Here, we report on the use of this bioinformatics program in identifying, classifying, and analyzing members of the HRGP superfamily (i.e. AGPs, EXTs, PRPs, hybrid HRGPs, and chimeric HRGPs) in the genetic model plant Arabidopsis. An overview of this bioinformatics approach is presented in Figure 1. In addition, public databases and programs were accessed and utilized to extract relevant biological information on these HRGPs in terms of their expression patterns, most similar sequences via BLAST analysis, available genetic mutants, and coexpressed HRGP, glycosyl transferase (GT), prolyl 4-hydroxylase (P4H), and peroxidase genes in Arabidopsis. This information provides new insight to the HRGP superfamily and can be used by researchers to facilitate and guide further research in the field. Moreover, the bioinformatics tools developed here can be readily applied to protein sequences from other species to analyze their HRGPs or, for that matter, any given protein family by altering the input parameters.

RESULTS

Finding and Classifying AGPs

The BIO OHIO program was used to identify potential classical AGPs, including the Lys-rich classical AGPs, AG peptides, and chimeric AGPs (i.e. FLAs and other chimeric AGPs) from the Arabidopsis proteome (Table I). The program initially identified 64 possible classical AGPs by searching for biased amino acid compositions of at least 50% PAST. Similarly, 86 potential AG peptides were identified by searching for proteins between 50 and 90 amino acids in length with biased amino acid compositions of at least 35% PAST. Finally, 25 potential FLAs were identified by searching for the following fasciclin H1 motif: [MALIT]T[VILS][FLCM][CAVT][PVLIS][GSTKRNDPEIV]+[DNS][DSENAGE]+[ASQM]. The 175 proteins identified by the program were further examined individually to determine if they appeared to be AGPs. The presence of a signal peptide was one such factor, as was the presence and location of AP, PA, SP, and TP repeats, since these dipeptide sequences are often present in known AGPs (Nothnagel, 1997). Finally, the presence of a GPI anchor addition sequence provided additional support, although not all AGPs have this sequence. By these criteria, 64 of the original 175 were classified as AGPs; moreover, they fall into several distinct classes: 20 classical AGPs, three Lys-rich (classical) AGPs, 16 AG peptides, 21 chimeric FLAs, three chimeric plastocyanin AGPs (PAGs), and one other chimeric AGP (Tables I and II). Additionally, one other AGP was documented in the literature, AGP30, a nonclassical or chimeric AGP, but was not identified by the program given that its PAST value of 34% was below the 50% threshold value used by the program (Baldwin et al., 2001; van Hengel and Roberts, 2003). Consequently, this AGP was added to the list of AGPs appearing in Table II but was not counted in Table I. In addition, four PRPs (PRP18, PRP5, PRP6, PRP16), 20 EXTs (EXT40, EXT17, EXT38, EXT19, EXT22, EXT18, EXT15, EXT7, EXT9, EXT10, EXT2, EXT11, EXT13, EXT16, EXT6, EXT12, EXT14, EXT8, EXT20, EXT21), and three hybrid AGP/EXTs (HAEs; HAE1, HAE3, HAE4) were identified by the program using the 50% PAST rule; further information on these HRGP sequences is presented below.

Some AGPs, particularly chimeric AGPs, can be below the 50% PAST threshold but were identified by searching the Arabidopsis protein database annotations and then subjecting such proteins to further analysis (i.e. searching for signal peptides, AP, PA, SP, and TP repeats, or GPI anchor addition sequences). With this approach, 21 additional AGPs were found, including two classical AGPs (AGP50C and AGP57C), 14 PAGs, and five other chimeric AGPs, including AGP30. The locus identifiers of these sequences are indicated in italics in Table II.

With the addition of these AGPs from the protein database annotations, the total number of potential AGPs became 85 and included 22 classical AGPs, three Lys-rich classical AGPs, 16 AG peptides, 21 chimeric FLAs, 17 chimeric PAGs, and six other chimeric AGPs (Table II). Representative amino acid sequences of these potential AGPs, including the predicted locations of their signal peptides and GPI anchor addition sequences, are displayed in Figure 2 and Supplemental Figure S1. The classical AGPs ranged in size from 87 to 739 amino acids. The majority (19 of 22) were

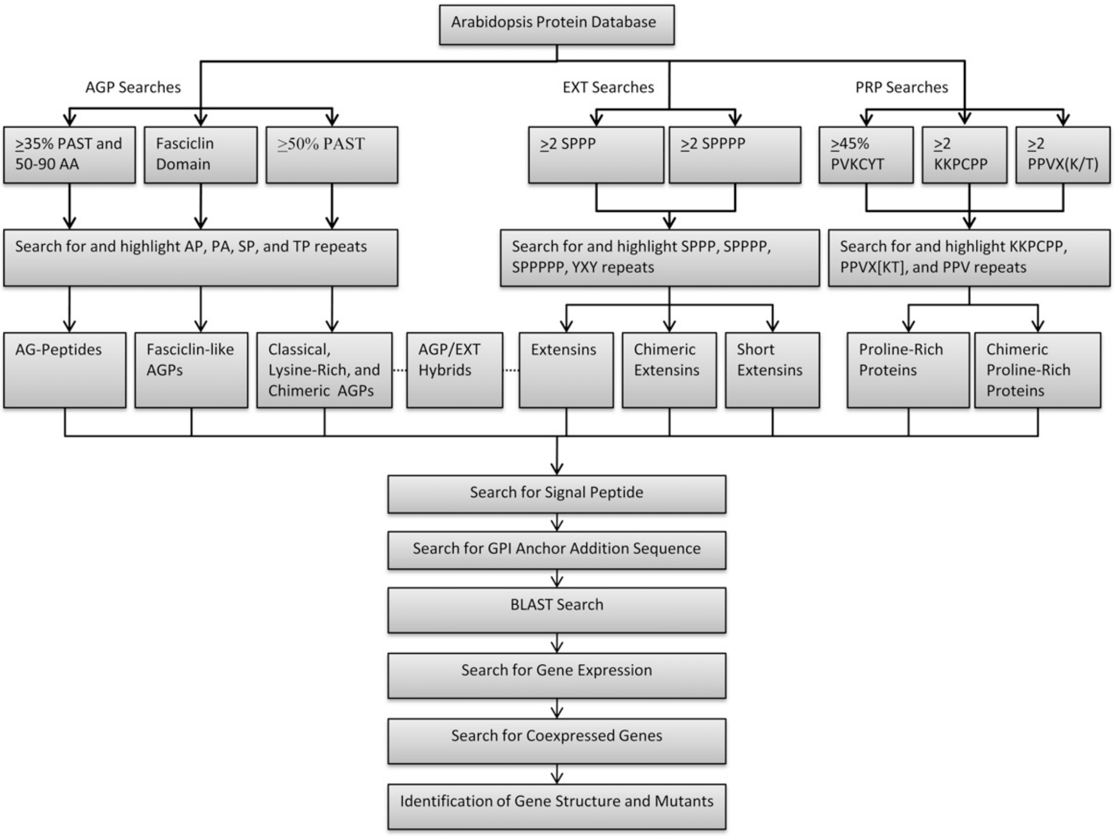

Arabidopsis Protein Database
AGP Searches
EXT Searches
PRP Searches
≥35% PAST and 50-90 AA
Fasciclin Domain
≥50% PAST
≥2 SPPP
≥2 SPPPP
≥45% PVKCYT
≥2 KKPCPP
≥2 PPVX(K/T)
Search for and highlight AP, PA, SP, and TP repeats
Search for and highlight SPPP, SPPPP, SPPPPP, YXY repeats
Search for and highlight KKPCPP, PPVX[KT], and PPV repeats
AG-Peptides
Fasciclin-like AGPs
Classical, Lysine-Rich, and Chimeric AGPs
AGP/EXT Hybrids
Extensins
Chimeric Extensins
Short Extensins
Proline-Rich Proteins
Chimeric Proline-Rich Proteins
Search for Signal Peptide
Search for GPI Anchor Addition Sequence
BLAST Search
Search for Gene Expression
Search for Coexpressed Genes
Identification of Gene Structure and Mutants

Figure 1. Bioinformatics workflow diagram summarizing the identification, classification, and analysis of HRGPs (AGPs, EXTs, and PRPs) in Arabidopsis. Classical AGPs were defined as containing greater than 50% PAST coupled with the presence of AP, PA, SP, and TP repeats distributed throughout the protein, Lys-rich AGPs were a subgroup of classical AGPs that included a Lys-rich domain, and chimeric AGPs were defined as containing greater than 50% PAST coupled with the localized distribution of AP, PA, SP, and TP repeats. AG peptides were defined to be 50 to 90 amino acids in length and containing greater than 35% PAST coupled with the presence of AP, PA, SP, and TP repeats distributed throughout the peptide. FLAs were defined as having a fasciclin domain coupled with the localized distribution of AP, PA, SP, and TP repeats. Extensins were defined as containing two or more SP_3 or SP_4 repeats coupled with the distribution of such repeats throughout the protein; chimeric extensins were similarly identified but were distinguished from the extensins by the localized distribution of such repeats in the protein; and short extensins were defined to be less than 200 amino acids in length coupled with the extensin definition. PRPs were identified as containing greater than 45% PVKCYT or two or more KKPCPP or PVX(K/T) repeats coupled with the distribution of such repeats and/or PPV throughout the protein. Chimeric PRPs were similarly identified but were distinguished from PRPs by the localized distribution of such repeats in the protein. Hybrid HRGPs (i.e. AGP/EXT hybrids) were defined as containing two or more repeat units used to identify AGPs, extensins, or PRPs. The presence of a signal peptide was used to provide added support for the identification of an HRGP but was not used in an absolute fashion. Similarly, the presence of a GPI anchor addition sequence was used to provide added support for the identification of classical AGPs and AG peptides, which are known to contain such sequences. BLAST searches were also used to provide some support to our classification if the query sequence showed similarity to other members of an HRGP subfamily. Note that some AGPs, particularly chimeric AGPs, and PRPs were identified from an Arabidopsis database annotation search and that two chimeric extensins were identified from the primary literature as noted in the text.

predicted to have a signal peptide, and many (14 of 22) were also predicted to have a GPI anchor. The Lys-rich, classical AGPs ranged in size from 185 to 247 amino acids. All three were predicted to have a signal peptide, but only two were predicted to have a GPI anchor. The AG peptides ranged in size from 58 to 87 amino acids. All 16 AG peptides were predicted to have a signal peptide, but only 12 were predicted to have a GPI anchor. The FLAs ranged in size from 247 to 462 amino acids. The majority (20 of 21) were predicted to have a signal peptide, but only 11 were predicted to have a GPI anchor. The FLAs are a type of chimeric AGP; each FLA contains either one or two AGP domains. Such AGP domains were readily visualized with the BIO OHIO program by utilizing the sliding windows feature to search for biased amino acid sequences within a user-defined amino acid window size (e.g. 80% PAST in a 10-amino acid window) that slides along the protein sequence. Usually, such domains were also apparent by examining the location of

Table I. *AGPs identified from the Arabidopsis genome based on biased amino acid compositions, size, and the presence of fasciclin domains*

The number in parentheses indicates the number of proteins that had a predicted signal peptide sequence.

Search Criteria	Total	Classical AGP	Lys-Rich AGP	AG Peptide	FLA	Chimeric AGP	PRP	EXT	Hybrid	Others
≥50% PAST	64 (47)	19 (16)	3 (3)	1 (1)	0	4 (4)	4 (3)	20 (17)	3 (2)	10 (1)
≥35% PAST and 50 to 90 amino acids	86 (36)	1 (1)	0	16 (16)	0	0	0	0	0	69 (20)
Fasciclin domain	25 (21)	0	0	0	21 (20)	0	0	0	0	4 (2)

the AP, PA, SP, and TP repeat units, which was easily done by the BIO OHIO program. The PAGs ranged in size from 177 to 370 amino acids. The 17 PAGs were all predicted to have a signal peptide, and 16 were predicted to have a GPI anchor. The other chimeric AGPs ranged in size from 222 to 826 amino acids. All but one (five of six) of these chimeric AGPs were predicted to have a signal peptide, and only one was predicted to have a GPI anchor as well as a signal peptide.

BLAST analysis was also conducted using The Arabidopsis Information Resource (TAIR) WU-Blast 2.0 to identify other potential AGP sequences and to provide insight to AGP sequences with the greatest similarity (Table II; Supplemental Table S1). BLAST searches were initially conducted with the filtering option on, but they were repeated with filtering off for those searches that found no other HRGPs. Such analysis showed that not all AGPs can be found with this method, but it did reveal sequences showing high degrees of similarity. BLAST was most successful for locating other FLAs and PAGs. In other words, a BLAST search using any one FLA sequence found most, but typically not all, other known FLA sequences.

AGP Gene Expression and Coexpressed HRGPs, GTs, P4Hs, and Peroxidases

In order to elucidate patterns of gene expression for these predicted AGPs, three public databases were searched: Genevestigator (https://www.genevestigator.ethz.ch/), the Arabidopsis Membrane Protein Library (http://www.cbs.umn.edu/arabidopsis/), and the Arabidopsis Massively Parallel Signature Sequencing (MPSS) Plus Database (http://mpss.udel.edu/at/). While about half of the AGPs had a broad range of expression throughout the plant, the other half showed organ-specific expression. Notably, several AGPs were specifically or preferentially expressed in the pollen, while others were expressed in roots, stems, leaves, and siliques (Table II; Supplemental Figs. S2–S5). Moreover, in examining the expression levels of all the AGP genes, the ones specifically or preferentially expressed in the pollen were the most highly expressed, as indicated by their high relative signal intensities. Furthermore, there was no observed correlation between organ-specific expression and a particular AGP class or between environmental stress-induced expression and a particular AGP class.

In order to elucidate HRGP gene networks and identify genes involved with AGP biosynthesis, the AGP genes were next examined with respect to coexpressed genes using The Arabidopsis Co-Response Database (http://csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html; Table III; Supplemental Table S2). Unfortunately, 39 of the 85 AGPs had no coexpression data available, so the following information was based on the 46 AGPs for which data were available. In analyzing the data, a focus was placed not only on other HRGPs but on GTs, P4Hs, and peroxidases, since GTs and P4Hs, and possibly peroxidases (Kjellbom et al., 1997), are responsible for posttranslational modification of AGPs. In terms of AGPs being expressed with other HRGPs, a total of 73 HRGPs were coexpressed with one or more AGPs. Among all HRGPs, FLA7 was coexpressed with the most AGPs, a total of 22 different AGPs. Interestingly, several different EXT and PRP genes were also coexpressed with numerous AGP genes. For the GTs, 27 of the 42 members of the GT2 family, 17 of the 42 members of the GT8 family, 11 of the 33 members of the GT47 family, and two of the three members of the GT29 family were coexpressed with various AGPs, to name just a few. Most notably, two members of the GT47 family (At5g22940 and At4g38040) were found to be coexpressed with 17 and 15 AGP genes, respectively. Also notable was the one member of the GT29 family (At1g08660) that was coexpressed with 14 different AGP genes and the three members of the GT8 family (At1g24170, At5g47780, At1g13250) that were coexpressed with 13, 11, and 10 different AGPs, respectively. In conducting this GT analysis, it was observed that not all of the CAZY members are annotated as GTs in the coexpression database. Consequently, coexpressed genes had to be cross-referenced against the gene identifiers listed in the CAZY database. For the P4Hs, five of 13 members of the P4H gene family were coexpressed with various AGPs. Among these, one P4H gene (At3g06300 or P4H2) was coexpressed with 10 different AGPs. Many peroxidase genes showed evidence of coexpression. The greatest amount of coexpression was exhibited by At4g26010, which was coexpressed with 13 different AGPs.

AGP Gene Organization and Mutants

Information was extracted from the TAIR and SALK Web sites with regard to the gene structure and avail-

Table II. *Identification, characterization, and classification of the AGP genes in Arabidopsis*

Locus Identifier[a]	Name[b]	Class	AP/PA/SP/TP Repeats	PAST	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top 5 BLAST Hit HRGPs[e]
At1g24520	**AGP50C**	Classical	4/1/3/1	43%	125	Yes	Yes	Pollen	0	1/0/1/0/0	AGP11C, AGP6C, PAG17, AGP10C, AGP4C
At1g31250	**AGP51C**	Classical	1/2/10/8	54%	165	Yes	No	Siliques	1	0/1/3/0/1	AGP9C, AGP58C, AGP33I, PRP18, EXT51
At1g35230	AGP5C	Classical	8/5/7/2	63%	133	Yes	Yes	Siliques, sepals	0	0/0/1/0/0	AGP10C, AGP7C, AGP4C, AGP2C, AGP1C
At1g63530	**AGP52C**	Classical	3/12/7/6	50%	499	No	No		1	1/1/3/3/0	AGP53C, AGP55C
At1g63540	**AGP53C**	Classical	9/15/21/7	51%	635	No	No	Pollen	1	0/0/2/0/3	AGP52C, AGP55C
At2g14890	AGP9C	Classical	9/11/13/7	68%	191	Yes	Yes		1	4/2/1/4/1	AGP18K, AGP17K, AGP15P, PAG13, PAG8
At2g22470	AGP2C	Classical	8/5/6/4	71%	131	Yes	Yes	Roots	0	2/6/0/0/6	AGP3C, AGP7C, AGP4C, AGP10C, AGP5C
At2g28440	**AGP54C**	Classical	5/5/28/0	63%	268	Yes	No	Pollen	0	2/0/3/0/0	AGP57C, AGP9C, AGP1C, HAE1, AGP11C
At2g45000	**AGP55C**	Classical	15/14/14/16	56%	739	No	No	Roots, pollen	8	0/6/7/4/4	AGP53C, AGP52C, LRX5, PAG10, PAG17
At2g47930	AGP26C	Classical	2/2/7/3	50%	136	Yes	Yes		0	4/1/0/0/0	HAE1, AGP2C, HAE4, PERK13
At3g01700	AGP11C	Classical	7/3/6/2	57%	136	Yes	Yes	Pollen	0	0/2/4/0/2	AGP6C, AGP21P
At3g06360	AGP27C	Classical	3/3/5/0	53%	125	Yes	Yes		0	6/2/4/0/2	AGP25C, AGP9C, AGP26C, AGP57C, AGP54C
At3g22070	**AGP56C**	Classical	4/3/7/3	61%	178	Yes	No		0	2/1/0/0/0	PERK8, LRX3, LRX5, EXT51, PEX3
At3g45230	**AGP57C**	Classical	1/3/16/0	53%	175	Yes	No		0	2/11/3/0/6	AGP54C
At4g09030	AGP10C	Classical	6/4/5/8	57%	127	Yes	Yes		0	1/2/1/0/3	AGP5C, AGP4C, AGP6C, AGP9C, AGP2C
At4g16980	**AGP58C**	Classical	3/1/8/4	42%	164	Yes	Yes		0	2/1/0/0/0	AGP50C
At4g40090	AGP3C	Classical	3/3/2/3	48%	87	Yes	No	Roots	0	4/0/1/0/1	AGP2C, PRP18
At5g10430	AGP4C	Classical	8/11/4/9	54%	135	Yes	Yes	Roots	0	3/2/2/0/0	AGP7C, AGP5C, PRP14, EXT32
At5g14380	AGP6C	Classical	9/3/8/1	48%	150	Yes	Yes	Pollen	0	2/0/0/0/0	AGP11C, AGP1C, AGP2C, FLA3, AGP9C
At5g18690	AGP25C	Classical	1/0/9/0	61%	116	Yes	Yes	Stems	0	7/7/1/0/3	AGP27C, AGP26C
At5g64310	AGP1C	Classical	7/8/12/1	72%	131	Yes	Yes	Roots	0	0/0/0/0/0	AGP7C, AGP2C, AGP18K, AGP4C, AGP15P
At5g65390	AGP7C	Classical	9/6/6/5	64%	130	Yes	Yes	Roots	0	2/0/1/0/3	AGP4C, AGP2C, AGP3C
At1g68725	AGP19K	Lys-rich	19/19/16/5	50%	247	Yes	No		1	0/0/1/0/0	AGP20P, AGP16P, AGP41P, AGP15P, AGP22P
At2g23130	AGP17K	Lys-rich	13/12/10/5	59%	185	Yes	Yes		1	1/0/12/0/0	AGP18K, AGP9C, AGP15P
At4g37450	AGP18K	Lys-rich	13/11/16/3	66%	209	Yes	Yes		1	6/2/3/0/1	AGP17K, AGP9C
At1g51915	**AGP42P**	AG peptide	2/1/1/0	35%	67	Yes	No	Stamen	1	0/0/1/0/0	None
At1g55330	AGP21P	AG peptide	2/2/1/0	46%	58	Yes	Yes		0	0/1/0/0/0	AGP12P, AGP13P, AGP14P
At2g41905	**AGP43P**	AG peptide	2/3/0/0	44%	61	Yes	Yes	nr[f]	0	2/3/1/0/2	AGP23P, PERK13
At2g46330	AGP16P	AG peptide	3/2/0/0	41%	73	Yes	No[g]		1	2/1/0/0/0	AGP20P, AGP41P, AGP22P, AGP15P, AGP21P
At3g01730	**AGP44P**	AG peptide	1/0/2/1	45%	87	Yes	Yes	Roots	0	1/0/3/0/3	AGP16P, EXT38
At3g13520	AGP12P	AG peptide	2/2/1/0	43%	60	Yes	Yes		0	0/0/0/0/1	AGP21P, AGP14P, AGP13P, AGP15P, AGP40P
At3g20865	**AGP40P**	AG peptide	1/1/2/0	48%	61	Yes	Yes	Pollen	0	4/1/2/0/2	AGP2C, AGP15P
At3g57690	AGP23P	AG peptide	2/3/0/0	45%	60	Yes	Yes	Pollen	0	6/1/0/0/4	AGP43P
At3g61640	AGP20P	AG peptide	2/1/2/0	43%	74	Yes	No		1	4/3/1/0/0	AGP16P, AGP41P, AGP22P, AGP15P, PAG6
At4g26320	AGP13P	AG peptide	2/2/1/0	47%	59	Yes	Yes	Roots	0	2/0/1/0/0	AGP14P, AGP12P, AGP21P
At5g11740	AGP15P	AG peptide	2/1/1/0	50%	61	Yes	Yes		0	2/4/0/0/1	AGP12P, AGP13P, AGP21P, AGP41P, AGP20P
At5g12880	**AGP45P**	AG peptide	1/0/3/0	43%	73	Yes	No	Roots	0	6/2/2/0/3	EXT17, EXT13, EXT20, EXT22, EXT15
At5g24105	**AGP41P**	AG peptide	3/2/0/0	38%	63	Yes	Yes	nr	1	3/2/0/1/0	AGP16P, AGP20P, AGP22P
At5g40730	AGP24P	AG peptide	3/3/0/0	40%	69	Yes	Yes	Pollen	0	3/0/0/0/1	PRP8
At5g53250	AGP22P	AG peptide	2/2/1/0	38%	63	Yes	Yes	Pollen, roots	1	1/0/0/0/1	AGP20P, AGP41P, AGP16P

(*Table continues on following page.*)

Table II. (*Continued from previous page.*)

Locus Identifier[a]	Name[b]	Class	AP/PA/SP/TP Repeats	PAST	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top 5 BLAST Hit HRGPs[e]
At5g56540	AGP14P	AG peptide	2/1/1/0	41%	60	Yes	Yes	Roots	0	3/4/2/0/1	AGP13P, AGP12P, AGP21P, EXT31, PAG9
At1g03870	FLA9	Chimeric	6/4/4/0	31%	247	Yes	Yes	Roots	0	4/2/3/0/2	FLA13, FLA6, FLA11, FLA12, FLA7
At1g15190	FLA19	Chimeric	3/2/7/0	33%	248	Yes	No		0	2/0/1/0/2	FLA21, FLA20, LRX3, HAE1, EXT18
At2g04780	FLA7	Chimeric	9/7/3/1	39%	254	Yes	Yes		1	4/0/1/0/2	FLA12, FLA9, FLA6, FLA13, FLA11
At2g20520	FLA6	Chimeric	5/3/2/1	34%	247	Yes	No	Roots	0	0/2/4/0/1	FLA9, FLA13, FLA11, FLA12, FLA7
At2g24450	FLA3	Chimeric	11/7/4/2	38%	280	Yes	Yes	Pollen	0	0/2/0/0/1	FLA5, FLA14, FLA10, FLA8, FLA2
At2g35860	FLA16	Chimeric	9/6/3/1	28%	445	Yes	No		1	1/1/1/2/3	FLA15, FLA17, FLA18, FLA12, FLA13
At2g45470	FLA8	Chimeric	13/6/8/3	43%	420	Yes	Yes		0	4/2/5/0/1	FLA10, FLA1, FLA2, FLA14, FLA3
At3g11700	FLA18	Chimeric	8/3/1/0	25%	462	Yes	No		2	8/3/7/5/0	FLA17, FLA15, FLA16, FLA6, FLA12
At3g12660	FLA14	Chimeric	2/2/4/0	35%	255	Yes	Yes	Stamen	0	2/2/0/0/0	FLA10, FLA8, FLA3, FLA1, FLA2
At3g46550	FLA4	Chimeric	1/4/4/1	37%	420	Yes	No		0	3/3/4/0/0	FLA10, FLA12, FLA6, FLA9, FLA11
At3g52370	FLA15	Chimeric	10/4/2/1	28%	436	Yes	No	Roots	1	5/6/6/1/0	FLA16, FLA18, FLA17, FLA12, FLA6
At3g60900	FLA10	Chimeric	13/7/7/4	41%	422	Yes	Yes	Siliques, carpel	0	10/8/5/0/3	FLA8, FLA1, FLA2, FLA14, FLA3
At4g12730	FLA2	Chimeric	4/2/3/0	31%	403	Yes	No		0	1/0/1/0/1	FLA1, FLA8, FLA10, FLA14, FLA3
At4g31370	FLA5	Chimeric	6/6/3/3	37%	278	Yes	Yes		0	1/0/3/0/0	FLA3, FLA14, FLA10, FLA2, FLA8
At5g03170	FLA11	Chimeric	6/3/0/0	36%	246	Yes	Yes	Stems	0	2/0/6/0/0	FLA12, FLA9, FLA13, FLA6, FLA7
At5g06390	FLA17	Chimeric	9/5/2/0	26%	458	Yes	No		2	12/2/6/1/0	FLA18, FLA15, FLA16, FLA12, FLA13
At5g06920	FLA21	Chimeric	0/0/6/2	32%	353	Yes	No		0	0/2/4/0/0	FLA19, FLA20
At5g40940	FLA20	Chimeric	2/0/4/1	29%	424	No	No		0	0/0/3/0/1	FLA21, FLA19, FLA12
At5g44130	FLA13	Chimeric	5/2/4/1	30%	247	Yes	Yes		0	2/1/0/0/0	FLA9, FLA6, FLA11, FLA12, FLA7
At5g55730	FLA1	Chimeric	9/6/3/1	33%	424	Yes	Yes		1	5/0/4/1/0	FLA2, FLA8, FLA10, FLA14, FLA3
At5g60490	FLA12	Chimeric	6/6/2/1	35%	249	Yes	Yes	Stems	0	9/0/1/0/0	FLA11, FLA13, FLA9, FLA6, FLA7
At2g23990	**PAG1**	Chimeric	7/7/3/3	39%	207	Yes	Yes		1	0/1/0/0/1	PAG12, PAG2, PAG15, PAG13, PAG7
At2g25060	**PAG2**	Chimeric	3/3/3/0	31%	182	Yes	Yes		1	3/3/2/1/0	PAG13, PAG15, PAG12, PAG1, PAG7
At2g26720	**PAG3**	Chimeric	1/2/3/1	30%	206	Yes	Yes		0	2/0/0/0/0	PAG4, PAG16, PAG5, PAG8, At3g53330
At2g31050	**PAG4**	Chimeric	3/2/4/0	32%	200	Yes	Yes	Pollen	0	1/0/0/0/1	PAG3, PAG16, PAG5, PAG8, At3g53330
At2g32300	**PAG5**	Chimeric	3/4/6/2	46%	261	Yes	Yes	Roots	2	0/0/1/0/0	PAG3, PAG4, PAG16, PAG8, PAG2
At2g44790	**PAG6**	Chimeric	0/1/3/9	42%	202	Yes	Yes	Roots	1	1/0/3/1/4	PAG9, PAG8, PAG5, PAG3, PAG4
At3g20570	**PAG7**	Chimeric	4/3/4/3	38%	203	Yes	Yes		1	5/1/1/0/1	PAG2, PAG15, PAG13, PAG12, PAG17
At3g60270	**PAG8**	Chimeric	3/1/8/1	38%	187	Yes	Yes	Roots	1	8/0/2/0/0	PAG9, PAG6, PAG4, PAG3, PAG16
At3g60280	**PAG9**	Chimeric	2/2/9/7	50%	222	Yes	Yes	Roots	1	1/0/0/0/5	PAG8, PAG6, PAG3, PAG5, PAG16

(*Table continues on following page.*)

Table II. (*Continued from previous page.*)

Locus Identifier[a]	Name[b]	Class	AP/PA/SP/TP Repeats	PAST	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top 5 BLAST Hit HRGPs[e]
At4g27520	**PAG10**	Chimeric	10/4/20/4	52%	349	Yes	Yes		1	7/4/0/0/0	PAG17, PAG14, PAG11, PAG2, PAG7
At4g28365	**PAG11**	Chimeric	2/2/6/1	31%	199	Yes	Yes		1	4/2/1/0/0	PAG14, PAG10, PAG17, PAG12, PAG7
At4g30590	**PAG12**	Chimeric	4/3/3/1	31%	190	Yes	Yes		1	4/3/1/0/0	PAG1, PAG15, PAG13, PAG2, PAG7
At4g31840	**PAG13**	Chimeric	1/1/3/1	31%	177	Yes	Yes		1	0/1/7/0/5	PAG2, PAG15, PAG12, PAG1, PAG7
At4g32490	**PAG14**	Chimeric	5/4/6/3	33%	221	Yes	Yes	Siliques	1	1/5/1/0/3	PAG11, PAG10, PAG17, PAG2, PAG15
At5g25090	**PAG15**	Chimeric	3/4/4/0	32%	186	Yes	Yes	Shoot apex	1	5/2/3/1/3	PAG2, PAG12, PAG13, PAG7, PAG1
At5g26330	**PAG16**	Chimeric	0/2/2/1	40%	187	Yes	No		1	0/0/1/1/3	PAG3, PAG4, PAG5, PAG8, At3g53330
At5g53870	**PAG17**	Chimeric	10/15/32/9	54%	370	Yes	Yes		1	6/4/1/0/8	PAG10, PAG11, PAG14, PAG7, PAG1
At1g03820	**AGP28I**	Chimeric	2/2/1/1	24%	222	Yes	No		0	8/0/1/0/4	PAG7
At1g28290	**AGP31I**	Chimeric	10/6/5/2	43%	359	Yes	No	Roots	1	1/0/7/1/2	AGP30I, PRP1, PRP11, PRP7, PAG17
At1g36150	**AGP29I**	Chimeric	1/4/20/4	54%	256	Yes	Yes	Stamen	2	2/0/1/0/0	PEX1, PEX3, PERK8, HAE1, AGP19K
At2g33790	**AGP30I**	Chimeric	4/4/1/0	34%	239	Yes	No	Roots	1	7/0/1/0/0	AGP31I, PRP7, PRP11, PRP3, PRP1
At5g21160	**AGP32I**	Chimeric	8/8/9/2	30%	826	No	No		14	1/3/7/9/3	LRX5, LRX3, PEX1, PEX3, LRX2
At5g56330	**AGP33I**	Chimeric	18/18/2/10	39%	350	Yes	No	Stamen	6	1/2/2/3/1	EXT51, LRX3, PRP16, PEX4, PRP17

[a] Italics indicate a protein found using the Arabidopsis database annotation search. [b] Boldface indicates a protein that was not previously identified by Schultz et al. (2002). The letter designations in the names represent the following: C, classical AGP; P, AG peptide; K, Lys-rich classical AGP; I, chimeric AGP. [c] Signal peptide. [d] Indicates the number of mutants available in each location: P, promoter; 5, 5′ UTR; E, exon; I, intron; 3, 3′ UTR. [e] Underline indicates the result of a BLAST search with filtering turned off. [f] nr, Not reported. This indicates that data for a particular protein are not found in Genevestigator, Arabidopsis Membrane Protein Library, or MPSS. [g] Experimentally found to be GPI anchored (Schultz et al., 2004).

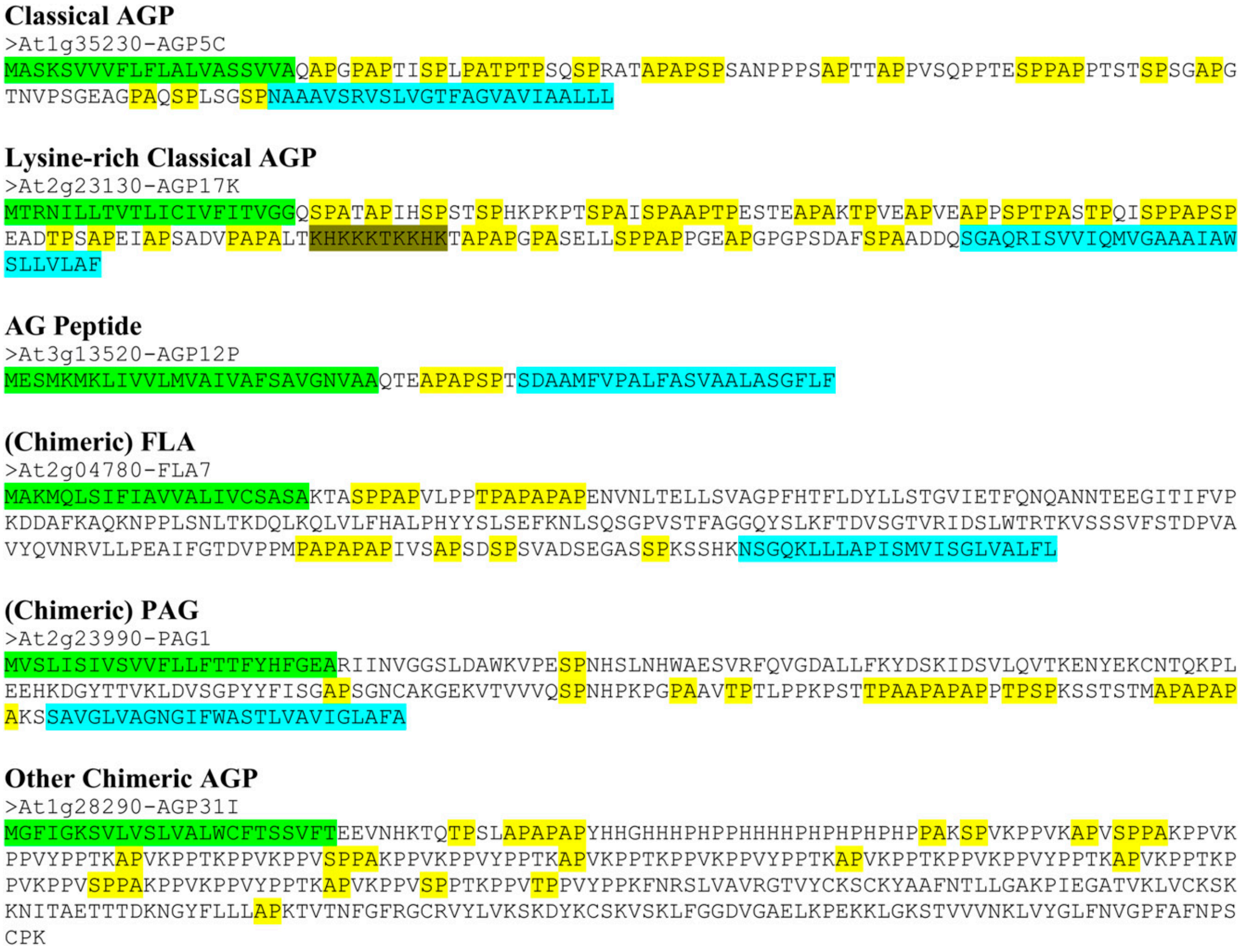
able genetic mutants for each of the predicted AGP genes. The AGP genes contained few, if any, introns. Of the 85 AGPs, 46 had no introns and 32 had only one intron (Table II; Supplemental Table S3). One chimeric AGP (At5g21160 or AGP32I), however, was predicted to have 14 introns.

Examination of the various mutant lines available for research showed that nearly 99% (84 of 85) of the AGP genes had one or more mutants available. Of these mutants, 33% were in the promoter region, 19% were in the 5′ untranslated region (UTR), 25% were in an exon, 6% were in an intron, and 17% were in the 3′ UTR (Table II; Supplemental Table S4).

Finding and Classifying EXTs

The BIO OHIO program was used to identify potential EXTs by searching for SP$_3$ and SP$_4$ sequences repeated two or more times (Table IV). The program initially identified 114 and 63 potential EXTs by searching for these tetrapeptide and pentapeptide repeats, respectively.

The 114 and 63 proteins identified by the program were further examined individually to determine if they appeared to be EXTs, with the realization that the 63 proteins are a subset of the 114. The presence of a signal peptide was one such factor, as was the presence and location of SP$_3$, SP$_4$, and SP$_5$ repeats, since these peptide sequences are often present in known EXTs. GPI anchor addition sequences are not known to be associated with EXTs; nonetheless, testing for the presence of such a sequence was performed out of curiosity. By these criteria, 57 of the 114 and 50 of the 63 proteins were classified as EXTs. While the SP$_4$ criteria resulted in a high percentage of EXT sequences, they did not locate all potential EXTs, given that the SP$_3$ criteria were used to find more EXTs, but with a higher rate of false positives. Subsequent analysis involved examining the 57 EXT sequences and attempting to classify them. Based upon the repeat sequences found in these EXTs, they were placed into nine classes: three SP$_5$ EXTs, two SP$_5$/SP$_4$ EXTs, 12 SP$_4$ EXTs, two SP$_4$/SP$_3$ EXTs, one SP$_3$ EXT, 12 short EXTs, 11 (chimeric) Leu-rich repeat EXTs (LRXs) that include pollen



Classical AGP

```
>At1g35230-AGP5C
MASKSVVVFLFLALVASSVVAQAPGPAPTISPLPATPTPSQSPRATAPAPSPSANPPPSAPTTAPPVSQPPTESPPAPPTSTSPSGAPG
TNVPSGEAGPAQSPLSGSPNAAAVSRVSLVGTFAGVAVIAALLL
```

Lysine-rich Classical AGP

```
>At2g23130-AGP17K
MTRNILLTVTLICIVFITVGGQSPATAPIHSPSTSPHKPKPTSPAISPAAPTPESTEAPAKTPVEAPVEAPPSPTPASTPQISPPAPSP
EADTPSAPEIAPSADVPAPALTKHKKKTKKHKTAPAPGPASELLSPPAPPGEAPGPGPSDAFSPAADDQSGAQRISVVIQMVGAAAIAW
SLLVLAF
```

AG Peptide

```
>At3g13520-AGP12P
MESMKMKLIVVLMVAIVAFSAVGNVAAQTEAPAPSPTSDAAMFVPALFASVAALASGFLF
```

(Chimeric) FLA

```
>At2g04780-FLA7
MAKMQLSIFIAVVALIVCSASAKTASPPAPVLPPTPAPAPAPENVNLTELLSVAGPFHTFLDYLLSTGVIETFQNQANNTEEGITIFVP
KDDAFKAQKNPPLSNLTKDQLKQLVLFHALPHYYSLSEFKNLSQSGPVSTFAGGQYSLKFTDVSGTVRIDSLWTRTKVSSSVFSTDPVA
VYQVNRVLLPEAIFGTDVPPMPAPAPAPIVSAPSDSPSVADSEGASSPKSSHKNSGQKLLLAPISMVISGLVALFL
```

(Chimeric) PAG

```
>At2g23990-PAG1
MVSLISIVSVVFLLFTTFYHFGEARIINVGGSLDAWKVPESPNHSLNHWAESVRFQVGDALLFKYDSKIDSVLQVTKENYEKCNTQKPL
EEHKDGYTTVKLDVSGPYYFISGAPSGNCAKGEKVTVVVQSPNHPKPGPAAVTPTLPPKPSTTPAAPAPAPPTPSPKSSTSTMAPAPAP
AKSSAVGLVAGNGIFWASTLVAVIGLAFA
```

Other Chimeric AGP

```
>At1g28290-AGP31I
MGFIGKSVLVSLVALWCFTSSVFTEEVNHKTQTPSLAPAPAPYHHGHHHPHPPHHHHPHPHPHPHPPAKSPVKPPVKAPVSPPAKPPVK
PPVYPPTKAPVKPPTKPPVKPPVSPPAKPPVKPPVYPPTKAPVKPPTKPPVKPPVYPPTKAPVKPPTKPPVKPPVYPPTKAPVKPPTKP
PVKPPVSPPAKPPVKPPVYPPTKAPVKPPVSPPTKPPVTPPVYPPKFNRSLVAVRGTVYCKSCKYAAFNTLLGAKPIEGATVKLVCKSK
KNITAETTTDKNGYFLLLAPKTVTNFGFRGCRVYLVKSKDYKCSKVSKLFGGDVGAELKPEKKLGKSTVVVNKLVYGLFNVGPFAFNPS
CPK
```

Figure 2. Protein sequences encoded by representative AGP gene classes in Arabidopsis. Colored sequences at the N and C termini indicate predicted signal peptide (green) and GPI anchor (light blue) addition sequences if present. AP, PA, SP, and TP repeats (yellow) and Lys-rich regions (olive) are also indicated.

extensin-like (PEX) proteins, 11 (chimeric) Pro-rich extensin-like receptor kinases (PERKs), and three other chimeric EXTs (Tables IV and V; Fig. 3). YXY repeats were observed in most of the EXT sequences. Such sequences are involved in cross-linking EXTs (Brady et al., 1996, 1998; Schnabelrauch et al., 1996; Held et al., 2004; Cannon et al., 2008). Forty of the 59 EXTs identified contain this YXY sequence. Although YVY is the most common repeat, YIY, YYY, and YAY repeats also occur less frequently. Interestingly, several EXTs have a YPY sequence immediately following the signal peptide.

The Arabidopsis protein database annotations were searched, but no additional EXTs were found beyond those already identified by the program. Additionally, four other PERKs were documented in the literature but were not identified by the program, because three (At5g24400 or PERK2, At1g68690 or PERK9, At4g32710 or PERK14) were not included in the Arabidopsis protein database and one (At1g52290 or PERK15) found in the database contained only one SPP. The PERK14 sequence was subsequently found on the TAIR Web site but lacked SP_3/SP_4 repeats. Nonetheless, PERK14 and PERK15, being members of the PERK family and having publicly available sequences, were added in italics to the list of EXTs appearing in Table V and subjected to subsequent analyses. PERK2 and PERK9 were described as pseudogenes on the TAIR Web site and had no sequences available. Thus, they were not added to the table or analyzed further. In addition, two AGPs (AGP9C, AGP19K) and four HAEs (HAE1, HAE2, HAE3, HAE4) were identified by the program using the SP_3 rule. Analysis of these AGP sequences was already presented in the AGP section above; however, the four hybrid HRGPs were considered here along with the EXT family members.

The three other chimeric EXTs were annotated in the Arabidopsis protein database as late embryogenesis abundant protein (EXT50), expressed protein (EXT51), and plastocyanin-like protein (EXT52). EXT50, EXT51,

Table III. *HRGPs, GTs, P4Hs, and peroxidases coexpressed with AGPs*

HRGP Locus Identifier	Name	No. of Coexpressed AGPs	GT Locus Identifier	Name	(Family)	No. of Coexpressed AGPs	GT Locus Identifier Continued	Name	(Family)	No. of Coexpressed AGPs	P4H Locus Identifier	Name	No. of Coexpressed AGPs	Peroxidases Locus Identifier	Name	No. of Coexpressed AGPs
At2g04780	FLA7	22	At5g22940	-	(GT47)	17	At1g05570	Gsl06	(GT48)	3	At3g06300	P4H2	10	At4g26010	ATP13a	13
At1g03870	FLA9	19	At4g38040	-	(GT47)	15	At1g06780	-	(GT8)	3	At5g18900	P4H4	4	At1g05240	-	12
At4g12730	FLA2	19	At4g39350	CesA02	(GT2)	15	At1g07240	-	(GT1)	3	At2g17720	P4H5	2	At1g30870	-	12
At4g16140	EXT37	17	At1g08660	-	(GT29)	14	At1g16570	-	(GT33)	3	At2g43080	P4H1	1	At3g49960	-	12
At2g45470	FLA8	16	At1g24170	-	(GT8)	13	At1g30530	-	(GT1)	3	At5g66060	P4H10	1	At5g17820	PER57	12
At5g60490	FLA12	16	At4g02500	-	(GT34)	12	At1g67880	-	(GT17)	3				At5g67400	PER73	12
At4g16980	AGP58C	13	At1g02730	CslD5	(GT2)	11	At1g73160	-	(GT4)	3				At3g28200	-	10
At5g10430	AGP4C	13	At5g05170	CesA03	(GT2)	11	At2g35650	CslA07	(GT2)	3				At1g05260	PER3	9
At1g28290	AGP31I	12	At5g47780	-	(GT8)	11	At3g27540	-	(GT17)	3				At2g43480	-	9
At3g20570	PAG7	12	At5g50420	-	(GT68)	11	At3g50740	-	(GT1)	3				At5g24070	-	9
At4g26320	AGP13P	12	At1g13250	-	(GT8)	10	At4g04970	Gsl01	(GT48)	3				At5g40150	-	9
At5g56540	AGP14P	12	At1g23480	CslA03	(GT2)	10	At4g31590	CslC05	(GT2)	3				At1g77490	tAPX	6
At3g19430	EXT51	11	At1g70090	-	(GT8)	10	At5g14850	-	(GT22)	3				At4g09010	-	6
At3g45230	AGP57C	11	At2g03220	FUT1	(GT37)	10	At5g15050	-	(GT14)	3				At4g21960	PER42	6
At4g37450	AGP18K	11	At3g18170	-	(GT61)	10	At5g38460	-	(GT57)	3				At2g18980	-	5
At5g55730	FLA1	11	At3g24040	-	(GT14)	10	At5g41460	-	(GT31)	3				At2g25080	GPX1	5
At1g62500	PRP14	10	At5g03760	CslA09	(GT2)	10	At1g14080	FUT6	(GT37)	2				At3g01190	PER27	5
At2g47930	AGP26C	10	At5g19690	-	(GT66)	10	At1g18580	-	(GT8)	2				At4g11290	-	5
At3g06750	EXT34	10	At1g34130	-	(GT66)	9	At1g21480	-	(GT47)	2				At4g33420	-	5
At3g13520	AGP12P	10	At1g74380	-	(GT34)	9	At1g27120	-	(GT31)	2				At4g35970	-	5
At3g62680	PRP3	10	At2g15370	FUT5	(GT37)	9	At1g60470	-	(GT8)	2				At2g22420	PER17	4
At4g31840	PAG13	10	At2g31750	-	(GT1)	9	At1g68020	-	(GT20)	2				At2g41480	-	4
At5g53250	AGP22P	10	At2g32620	CslB04	(GT2)	9	At1g68470	-	(GT47)	2				At5g39580	-	4
At5g65390	AGP7C	10	At3g28180	CslC04	(GT2)	9	At1g71220	-	(GT24)	2				At5g42180	PER64	4
At1g03820	AGP28I	9	At5g22740	CslA02	(GT2)	9	At1g73370	-	(GT4)	2				At2g18140	-	3
At1g55330	AGP21P	9	At1g19360	-	(GT77)	8	At1g78800	-	(GT4)	2				At2g43350	-	3
At1g70990	EXT33	9	At2g22900	-	(GT34)	8	At2g20370	-	(GT47)	2				At4g30170	-	3
At3g11700	FLA18	9	At3g25140	-	(GT8)	8	At2g28080	-	(GT1)	2				At2g37130	PER21	2
At4g27520	PAG10	9	At3g62720	-	(GT34)	8	At2g29750	-	(GT1)	2				At3g49120	-	2
At1g52290	PERK15	8	At4g15290	CslB05	(GT2)	8	At2g35100	-	(GT47)	2				At4g37530	-	2
At4g13340	LRX3	8	At5g64740	CesA06	(GT2)	8	At2g41640	-	(GT61)	2				At5g22410	-	2
At2g10940	PRP15	7	At1g16900	-	(GT22)	7	At3g21750	-	(GT1)	2				At1g71695	PER12	1
At2g33790	AGP30I	7	At1g27440	-	(GT47)	7	At3g46970	-	(GT35)	2				At2g35380	PER20	1
At3g54590	EXT2	7	At1g34270	-	(GT47)	7	At3g50760	-	(GT8)	2				At3g21770	PER30	1
At3g60900	FLA10	7	At1g71070	-	(GT14)	7	At4g09500	-	(GT1)	2				At3g63080	-	1
At4g09030	AGP10C	7	At2g37585	-	(GT14)	7	At4g18230	-	(GT1)	2				At4g08390	sAPX	1
At5g06630	EXT13	7	At3g03050	CslD3	(GT2)	7	At4g18240	-	(GT5)	2				At4g35000	APX3	1
At5g44130	FLA13	7	At4g00300	-	(GT31)	7	At4g24000	CslG2	(GT2)	2				At5g06730	-	1
At4g18670	LRX5	6	At4g02130	-	(GT8)	7	At4g26940	-	(GT31)	2				At5g66390	PER72	1
At5g06640	EXT14	6	At5g09870	CesA05	(GT2)	7	At5g01220	-	(GT4)	2						
At5g11740	AGP15P	6	At5g11110	-	(GT4)	7	At5g07720	-	(GT34)	2						
At5g25090	PAG15	6	At5g16190	CslA11	(GT2)	7	At5g16510	-	(GT75)	2						
At5g64310	AGP1C	6	At5g17420	CesA07	(GT2)	7	At5g20410	-	(GT28)	2						
At1g23720	EXT6	5	At5g39990	-	(GT14)	7	At5g66690	-	(GT1)	2						

(*Table continues on following page.*)

Table III. (*Continued from previous page.*)

HRGP Locus Identifier	Name	No. of Coexpressed AGPs	GT Locus Identifier	Name	(Family)	No. of Coexpressed AGPs	GT Locus Identifier Continued	Name	(Family)	No. of Coexpressed AGPs	P4H Locus Identifier	Name	No. of Coexpressed AGPs	Peroxidases Locus Identifier	Name	No. of Coexpressed AGPs
At2g22470	AGP2C	5	At5g61840	-	(GT47)	7	At1g06000	-	(GT1)	1						
At2g25060	PAG2	5	At1g08280	-	(GT29)	6	At1g06410	-	(GT20)	1						
At2g35860	FLA16	5	At1g19710	-	(GT4)	6	At1g11720	-	(GT5)	1						
At3g24480	LRX4	5	At1g74800	-	(GT31)	6	At1g12990	-	(GT17)	1						
At3g28550	EXT9	5	At3g02350	-	(GT8)	6	At1g20575	-	(GT2)	1						
At3g52370	FLA15	5	At3g56000	CslA14	(GT2)	6	At1g23870	-	(GT20)	1						
At5g18690	AGP25C	5	At4g01220	-	(GT77)	6	At1g24070	CslA10	(GT2)	1						
At5g21160	AGP32I	5	At4g17770	-	(GT20)	6	At1g24100	-	(GT1)	1						
At5g40730	AGP24P	5	At4g32410	CesA01	(GT2)	6	At1g28710	-	(GT77)	1						
At2g24980	EXT7	4	At5g05860	-	(GT1)	6	At1g43620	-	(GT1)	1						
At4g32710	PERK14	4	At5g15650	-	(GT75)	6	At1g50580	-	(GT1)	1						
At5g03170	FLA11	4	At5g44030	CesA04	(GT2)	6	At1g60140	-	(GT20)	1						
At5g49280	EXT41	4	At5g55500	-	(GT61)	6	At1g64910	-	(GT1)	1						
At1g23040	EXT31	3	At1g53290	-	(GT31)	5	At1g64920	-	(GT1)	1						
At3g22120	PRP16	3	At2g24630	CslC08	(GT2)	5	At1g75420	-	(GT4)	1						
At3g24550	PERK1	3	At2g35610	-	(GT77)	5	At1g77810	-	(GT31)	1						
At5g26330	PAG16	3	At2g44660	-	(GT57)	5	At2g15480	-	(GT1)	1						
At1g09460	PRP13	2	At3g05320	-	(GT65)	5	At2g19880	-	(GT21)	1						
At3g61640	AGP20P	2	At3g62660	-	(GT8)	5	At2g20810	-	(GT8)	1						
At5g09520	PRP9	2	At4g11350	-	(GT31)	5	At2g25300	-	(GT31)	1						
At5g14920	PRP18	2	At4g23490	-	(GT31)	5	At2g32430	-	(GT31)	1						
At1g26150	PERK10	1	At4g36890	-	(GT43)	5	At2g37090	-	(GT43)	1						
At2g21140	PRP2	1	At5g02410	-	(GT59)	5	At3g04240	-	(GT41)	1						
At2g43150	EXT8	1	At5g24300	-	(GT5)	5	At3g07330	CslC06	(GT2)	1						
At2g44790	PAG6	1	At5g62220	-	(GT47)	5	At3g11670	-	(GT4)	1						
At3g57690	AGP23P	1	At5g62620	-	(GT31)	5	At3g15940	-	(GT4)	1						
At4g08410	EXT10	1	At1g10400	-	(GT1)	4	At3g16520	-	(GT1)	1						
At4g30590	PAG12	1	At1g52420	-	(GT4)	4	At3g21790	-	(GT1)	1						
At5g15780	PRP11	1	At2g38650	-	(GT8)	4	At3g29630	-	(GT1)	1						
			At3g11420	-	(GT31)	4	At3g46720	-	(GT1)	1						
			At3g14570	Gsl04	(GT48)	4	At3g58790	-	(GT8)	1						
			At3g15350	-	(GT14)	4	At4g01070	-	(GT1)	1						
			At3g29320	-	(GT35)	4	At4g01750	-	(GT77)	1						
			At3g59100	Gsl11	(GT48)	4	At4g07960	CslC12	(GT2)	1						
			At3g61130	-	(GT8)	4	At4g15490	-	(GT1)	1						
			At4g31780	-	(GT28)	4	At4g18780	CesA08	(GT2)	1						
			At4g32120	-	(GT31)	4	At4g19460	-	(GT4)	1						
			At5g05890	-	(GT1)	4	At4g21060	-	(GT31)	1						
			At5g37180	-	(GT4)	4	At4g22580	-	(GT47)	1						
			At5g53340	-	(GT31)	4	At4g38240	-	(GT13)	1						
			At5g54690	-	(GT8)	4	At5g05900	-	(GT1)	1						
							At5g16910	CslD2	(GT2)	1						
							At5g44820	-	(GT77)	1						

Table IV. *EXTs identified from the Arabidopsis genome based on SP_3 and SP_4 amino acid repeat units*

The number in parentheses indicates the number of proteins that had a predicted signal peptide sequence.

Search Criteria	Total	EXT	AGP	PRP	Hybrid	Others
Two or more SP_3	114 (52)	57 (39)	2 (2)	0	4 (3)	51 (10)
Two or more SP_4	63 (41)	50 (36)	0	0	3 (2)	10 (3)

and EXT52 contained five, seven, and three SP_4 repeats, respectively. EXT51 also contained numerous TP and SP repeats, reminiscent of AGPs.

A hybrid HRGP was defined as a protein that contains sequence characteristics of different HRGPs, such as EXT and AGP sequence modules, within the same protein. The four hybrid proteins identified in the EXT search had sequence characteristics of both EXTs and AGPs. Three of these hybrids, HAE1, HAE3, and HAE4, were identified because they passed an EXT test as well as the classical AGP test, having at least 50% PAST and multiple PA and TP repeats. The other hybrid, HAE2, contained two SP_4 repeats and one additional SP_3 module but did not pass the 50% PAST threshold, having only 43% PAST. Nonetheless, it contained multiple AP, PA, SP, and TP repeats, which are indicative of AGPs.

BLAST analysis was also conducted with each of the EXTs, chimeric EXTs, and HAEs to identify other related sequences and to provide insight to EXT sequences with the greatest similarity (Table V; Supplemental Table S1). Such analysis showed that not all EXTs were found with this method but did reveal sequences showing high degrees of similarity and clearly showed many more potential EXT sequences compared with the results from the similar strategy for analysis of the AGPs. Such BLAST analysis of LRXs and PERKs proved especially effective, as a BLAST query using any one LRX or PERK resulted in the identification of all other members in their respective class. Analysis of the other chimeric EXTs revealed that only EXT52 resulted in BLAST hits; these hits were PAG17, PAG9, and PAG10. This result was expected, since EXT52 contains a plastocyanin domain along with the EXT motifs. BLAST analysis of the At4g11430 hybrid HRGP (HAE3) as the query sequence showed similarity to both AGP and EXT genes, providing support for its identification as a hybrid HRGP. BLAST results for the other HAEs were less informative, with HAE1 showing similarity to no other HRGPs and HAE2 and HAE4 showing similarity to only one PRP and multiple chimeric PRPs, respectively.

As seen in Table V and in Supplemental Figure S6, the 20 SP_5, SP_5/SP_4, SP_4, SP_4/SP_3, and SP_3 EXTs ranged in size from 212 to 1,018 amino acids. The majority (17 of 20) were predicted to have a signal peptide, and none was predicted to have a GPI anchor. The 12 short EXTs ranged in size from 96 to 181 amino acids. All but one was predicted to have a signal peptide, and surprisingly, seven were predicted to have a GPI anchor. The 11 LRXs ranged in size from 433 to 956 amino acids and consisted of an N-terminal Leu-rich repeat domain and a C-terminal EXT domain. All but two were predicted to have a signal peptide, and none was predicted to have a GPI anchor. The 13 PERKs ranged in size from 509 to 760 amino acids and consisted of an N-terminal EXT domain and a C-terminal kinase domain. None was predicted to have a signal peptide or a GPI anchor. The three chimeric EXTs contained three to seven diagnostic EXT repeats; two had signal peptides, and none contained GPI anchor addition sequences. The four HAEs contained 219 to 375 amino acids; three had a signal peptide and none had GPI anchor addition sequences. The EXT domains/motifs in the LRXs, PERKs, and other chimeric EXTs as well as the EXT/AGP hybrids were readily visualized with the BIO OHIO program by observing the locations of the SP_3, SP_4, and SP_5 repeat units.

EXT Gene Expression and Coexpressed HRGPs, GTs, P4Hs, and Peroxidases

In order to elucidate patterns of gene expression for these predicted EXTs, including the various chimeric EXTs and four HRGP hybrids, the same three public databases were searched as with the AGPs. While several EXTs had a broad range of expression throughout the plant, most of the EXT genes showed organ-specific expression. Notably, several EXTs were specifically or preferentially expressed in the root (27), while several others were specifically or preferentially expressed in the pollen/stamen (14) or siliques (one; Table V; Supplemental Figs. S7–S10). Moreover, in examining the expression levels of all the EXT genes, many of those specifically or preferentially expressed in the pollen were the most highly expressed ones, as indicated by their high relative signal intensities.

Next, the EXT and hybrid HRGP genes were examined with respect to coexpressed genes (Table VI; Supplemental Table S5). For EXTs, there was no information for 29 out of the 59 genes in The Arabidopsis Co-Response Database, and the four hybrid HRGP genes were also not listed in this database. In analyzing the data, a focus was placed not only on other HRGPs but on GTs, P4Hs, and peroxidases, since GTs, P4Hs, and EXT peroxidases are responsible for posttranslational modification of EXTs; this approach represents one potential avenue to identify genes involved in the posttranslational modification of EXTs. In terms of EXTs being expressed with other HRGPs, a total of 67 HRGPs were coexpressed with one or more EXTs. The most highly coexpressed HRGP was FLA2, which was coexpressed with a total of 15 EXTs, while

SP_5 EXT

```
>At1g26240-EXT20
MANPNGWPSLLMLVIALYSVSAHTSAQYTYSPPSPPSYVYKPPTHIYSSPPPPPYVYSSPPPPPYIYKSPPPPPYVYSSPPPPPYIYKS
PPPPPYVYSSPPPPPYIYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYNSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYS
SPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVY
KSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYNSPPPPPYVYKSPPPPPYV
YSSPPPSPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPYVYKSPPPPPYVYSSPPPPPY
VYKSPSPPPYVYKSPPPPPSYSYSYSSPPPPIY
```

SP_5/ SP_4 EXT

```
>At4g13390-EXT18
MISLRMKGLGHCLVYVVVFSVIAAIVTAYDSPSSTPQYTSPYPPKNYSPYLSESPPPPPPQYRRQEPKYTPHPEPNVYDSPTPLPYYFP
FPKLDIKSPPPPSVYTFSPPQLYYSPSPKVEYKSPPPPYVYSSLPPLTYYSPSPKVIYNSPPPPYIYSSPPPPPYYSPSPKVDYKSPPP
PYVYSSPPPPPYYSPSPKVEYKSPPPPYVYSFPPPPPYYSPSPKVGYKSPPAPYVYSSPPPPPYYSPSPKVNYKSPPPPYVYSSPPPPP
YSPSPKVEFKSPPPPYIYNSPPPPSYYSPSPKIDYKSPPPPYVYSSPPPPTYYSPSPRVDYKSPPPPYVYNSLPPPYVYNSPPPPPYYS
PSPTVNYKSPPPPYVYNSPPPPPYYSPFPKVEYKSPPPPYIYNSPPPPPYYSPSPKITYKSPPPPYIYKTPYY
```

SP_4 EXT

```
>At1g23720-EXT6
MVAASYEPYTYSSPPPPLYDSPTPKVDYKSPPPPYVYSSPPPPLSYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVEYKSPPPPYVYS
SPPPPYYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKPTYKSPPPPYVYNSPPPPYYSPSPKVEYKSPPPPYVYSSPPPPYYSPSPKVD
YKSPPPPYVYNSPPPPYYSPSPKPTYKSPPPPYIYSSPPPPYYSPSPKPVYKSPPPPYVYSSPPPPYYSPSPKPAYKSPPPPYVYSSPP
PPYYSPSPKPIYKSPPPPYVYNSPPPPYYSPSPKPAYKSPPPPYVYSFPPPPYYSPSPKPVYKSPPPPYVYNSPPPPYYSPSPKPAYKS
PPPPYVYSSPPPPYYSPSPKPTYKSPPPPYVYSSPPPPYYSPSPKPVYKSPPPPYIYNSPPPPYYSPSPKPSYKSPPPPYVYSSPPPPY
YSPSPKLTYKSSPPPYVYSSPPPPYYSPSPKVVYKSPPPPYVYSSPPPPYYSPSPKPSYKSPPPPYVYNSPPPPYYSPSPKVIYKSPPH
PHVCVCPPPPPCYSHSPKIEYKSPPTPYVYHSPPPPPYYSPSPKPAYKSSPPPYVYSSPPPPYYSPAPKPVYKSPPPPYVYNSPPPPYY
SPSPKPTYKSPPPPYVYSSPPPPYYSPTPKPTYKSPPPPYVYSSPPPPYYSPSPKPTYKSPPPPYVYSSPPPPYYSPAPKPTYKSPPPP
YVYSSPPPPYYSPSPKPTYKSPPPPYVYSSPPPPPYYSPSPKVEYKSPPPPYVYSSPPPPYYSPSPKVEYKSPPPPYVYSSPPPPPYYS
PSPKVEYKSPPPPYVYSSPPPPTYYSPSPKVEYKSPPPPYVYNSPPPPAYYSPSPKIEYKSPPPPYVYSSPPPPSYSPSPKAEYKSPPP
PSLYY
```

SP_4/ SP_3 EXT

```
>At1g21310-EXT3/5
MGSPMASLVATLLVLTISLTFVSQSTANYFYSSPPPPVKHYTPPVKHYSPPPVYHSPPPPKKHYEYKSPPPPVKHYSPPPVYHSPPPPK
KHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVY
KSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPP
PVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKKHYVYKSPPPPVKHY
SPPPVYHSPPPPKKHYVYKSPPPPVKHYSPPPVYHSPPPPKEKYVYKSPPPPPVHHYSPPHHPYLYKSPPPPYHY
```

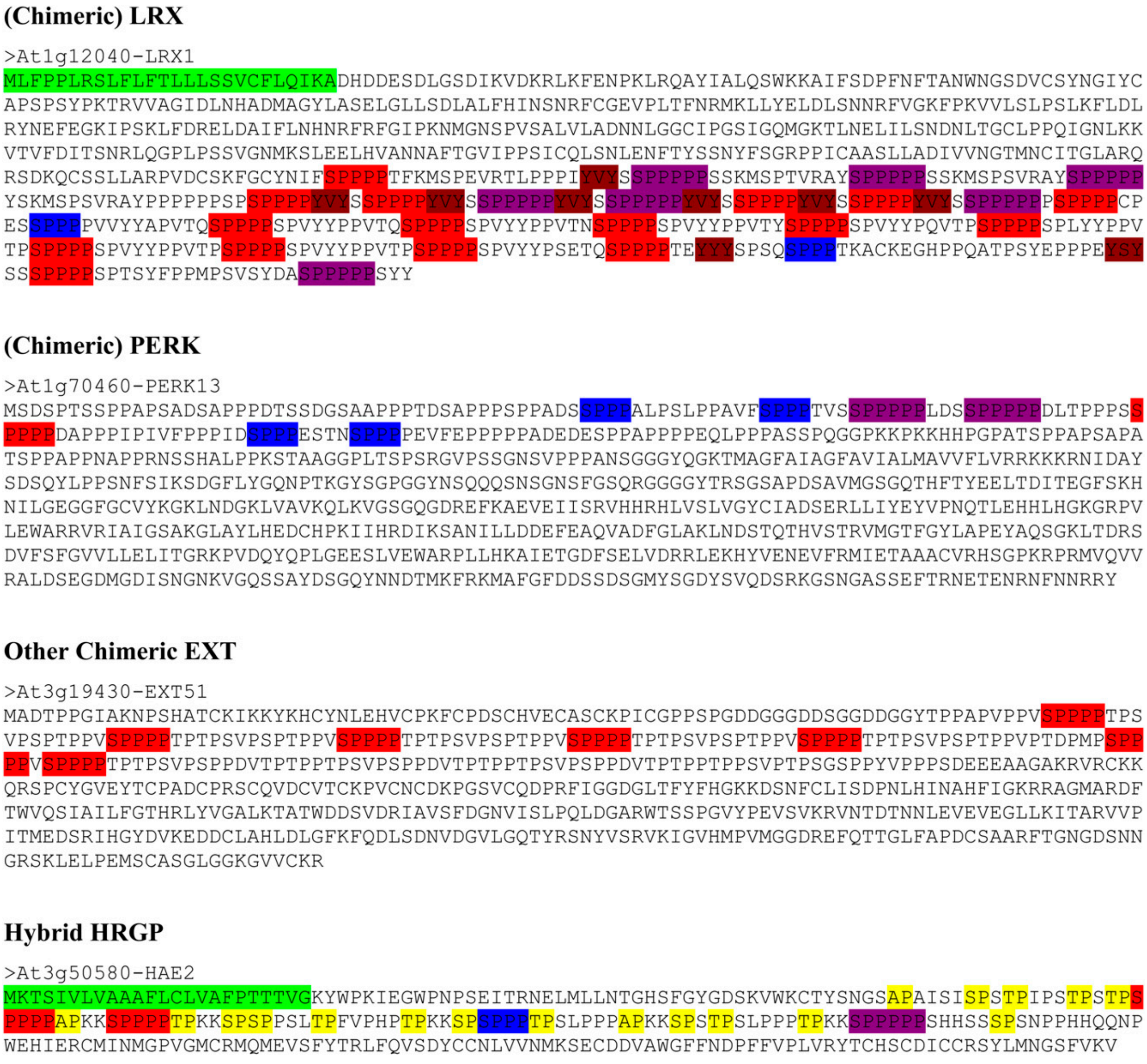
SP_3 EXT

```
>At4g08380-EXT17
MANPSNWPSLLMVILALYAVAAHTSAQYPYSPPSPPPYVYSSPPPYTYSPPPSPYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYS
SPPPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYSSPPPYVYSSPPPYAYSPPPYAYSPPPSPYVYKSPPYVYSSP
PPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPS
PYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYSSPPPYAYSPPPSPYVYKSPPYVYSSPPPY
AYSPPPSPYVYKSPPYVYSSPPPYTYSPPPYAYSPPPPCPDVYKPPPYVYSSPPPYVYNPPPSSPPPSPSYSYSSPPPPIY
```

Short EXT

```
>At4g16140-EXT37
METFRTFHLFLFFFFFTFTTTLTSPSQIADCTMCTSCDNPCQPNPSPPPPPSNPSPPPPSPTTTACPPPPSSSGGGPYYYYPPASQSGS
YRPPPSSSSGGYYYPPPKSGGNYPYTPPPNPIVPYFPFYYYNPPPQSVMSGSDAKIRFSYGVSFILIFSLYFGCF
```

Figure 3. (*Figure continues on following page.*)



(Chimeric) LRX

```
>At1g12040-LRX1
MLFPPLRSLFLFTLLLSSVCFLQIKADHDDESDLGSDIKVDKRLKFENPKLRQAYIALQSWKKAIFSDPFNFTANWNGSDVCSYNGIYC
APSPSYPKTRVVAGIDLNHADMAGYLASELGLLSDLALFHINSNRFCGEVPLTFNRMKLLYELDLSNNRFVGKFPKVVLSLPSLKFLDL
RYNEFEGKIPSKLFDRELDAIFLNHNRFRFGIPKNMGNSPVSALVLADNNLGGCIPGSIGQMGKTLNELILSNDNLTGCLPPQIGNLKK
VTVFDITSNRLQGPLPSSVGNMKSLEELHVANNAFTGVIPPSICQLSNLENFTYSSNYFSGRPPICAASLLADIVVNGTMNCITGLARQ
RSDKQCSSLLARPVDCSKFGCYNIFSPPPPTFKMSPEVRTLPPPIYVYSSPPPPPSSKMSPTVRAYSPPPPPSSKMSPSVRAYSPPPPP
YSKMSPSVRAYPPPPPPSPSPPPPYVYSSPPPPYVYSSPPPPPYVYSSPPPPPYVYSSPPPPYVYSSPPPPYVYSSPPPPPPSPPPPCP
ESSPPPPVVYYAPVTQSPPPPSPVYYPPVTQSPPPPSPVYYPPVTNSPPPPSPVYYPPVTYSPPPPSPVYYPQVTPSPPPPSPLYYPPV
TPSPPPPSPVYYPPVTPSPPPPSPVYYPPVTPSPPPPSPVYYPSETQSPPPPTEYYYSPSQSPPPTKACKEGHPPQATPSYEPPPEYSY
SSSPPPPSPTSYFPPMPSVSYDASPPPPPSYY
```

(Chimeric) PERK

```
>At1g70460-PERK13
MSDSPTSSPPAPSADSAPPPDTSSDGSAAPPPTDSAPPPSPPADSSPPPALPSLPPAVFSPPPTVSSPPPPPLDSSPPPPPDLTPPPSS
PPPPDAPPPIPIVFPPPIDSPPPESTNSPPPPEVFEPPPPPADEDESPPAPPPPEQLPPPASSPQGGPKKPKKHHPGPATSPPAPSAPA
TSPPAPPNAPPRNSSHALPPKSTAAGGPLTSPSRGVPSSGNSVPPPANSGGGYQGKTMAGFAIAGFAVIALMAVVFLVRRKKKRNIDAY
SDSQYLPPSNFSIKSDGFLYGQNPTKGYSGPGGYNSQQQSNSGNSFGSQRGGGGYTRSGSAPDSAVMGSGQTHFTYEELTDITEGFSKH
NILGEGGFGCVYKGKLNDGKLVAVKQLKVGSGQGDREFKAEVEIISRVHHRHLVSLVGYCIADSERLLIYEYVPNQTLEHHLHGKGRPV
LEWARRVRIAIGSAKGLAYLHEDCHPKIIHRDIKSANILLDDEFEAQVADFGLAKLNDSTQTHVSTRVMGTFGYLAPEYAQSGKLTDRS
DVFSFGVVLLELITGRKPVDQYQPLGEESLVEWARPLLHKAIETGDFSELVDRRLEKHYVENEVFRMIETAAACVRHSGPKRPRMVQVV
RALDSEGDMGDISNGNKVGQSSAYDSGQYNNDTMKFRKMAFGFDDSSDSGMYSGDYSVQDSRKGSNGASSEFTRNETENRNFNNRRY
```

Other Chimeric EXT

```
>At3g19430-EXT51
MADTPPGIAKNPSHATCKIKKYKHCYNLEHVCPKFCPDSCHVECASCKPICGPPSPGDDGGGDDSGGDDGGYTPPAPVPPVSPPPPTPS
VPSPTPPVSPPPPTPTPSVPSPTPPVSPPPPTPTPSVPSPTPPVSPPPPTPTPSVPSPTPPVSPPPPTPTPSVPSPTPPVPTDPMPSPP
PPVSPPPPTPTPSVPSPPDVTPTPPTPSVPSPPDVTPTPPTPSVPSPPDVTPTPPTPPSVPTPSGSPPYVPPPSDEEEAAGAKRVRCKK
QRSPCYGVEYTCPADCPRSCQVDCVTCKPVCNCDKPGSVCQDPRFIGGDGLTFYFHGKKDSNFCLISDPNLHINAHFIGKRRAGMARDF
TWVQSIAILFGTHRLYVGALKTATWDDSVDRIAVSFDGNVISLPQLDGARWTSSPGVYPEVSVKRVNTDTNNLEVEVEGLLKITARVVP
ITMEDSRIHGYDVKEDDCLAHLDLGFKFQDLSDNVDGVLGQTYRSNYVSRVKIGVHMPVMGGDREFQTTGLFAPDCSAARFTGNGDSNN
GRSKLELPEMSCASGLGGKGVVCKR
```

Hybrid HRGP

```
>At3g50580-HAE2
MKTSIVLVAAAFLCLVAFPTTTVGKYWPKIEGWPNPSEITRNELMLLNTGHSFGYGDSKVWKCTYSNGSAPAISISPSTPIPSTPSTPS
PPPPAPKKSPPPPTPKKSPSPPSLTPFVPHPTPKKSPSPPPTPSLPPPAPKKSPSTPSLPPPTPKKSPPPPPSHHSSSPSNPPHHQQNP
WEHIERCMINMGPVGMCRMQMEVSFYTRLFQVSDYCCNLVVNMKSECDDVAWGFFNDPFFVPLVRYTCHSCDICCRSYLMNGSFVKV
```

Figure 3. Protein sequences encoded by representative EXT and hybrid HRGP gene classes in Arabidopsis. Colored sequences at the N and C termini indicate predicted signal peptide (green) and GPI anchor (light blue) addition sequences if present. SP_3 (blue), SP_4 (red), SP_5 (purple), and YXY (dark red) repeats are also indicated. AP, PA, SP, and TP (yellow) repeats are indicted on hybrid HRGP only.

FLA9 was next on the list, being coexpressed with 14 EXTs. As reported above, FLA2 and FLA9 were also coexpressed with many AGP genes. A number of EXT genes, including EXT9, EXT13, EXT14, EXT6, EXT10, EXT2, and LRX4, were also coexpressed with 10 or more EXT genes.

For the GTs, the most coexpressed was CslB04, a member of the GT2 family, which was coexpressed with nine EXTs. Also highly coexpressed were At1g24170 (GT8), At1g74380 (GT34), At4g15290 (GT2), and At5g22940 (GT47), all of which were coexpressed with seven EXTs. Notably, several of the GTs that were coexpressed with EXTs were also coexpressed with AGPs. For example, one member of the GT8 family, At1g24170, was coexpressed with seven different EXTs and 13 different AGPs. For the P4Hs, four of 13 members of the P4H gene family were coexpressed with various EXTs. Among these, one P4H gene (At3g06300 or P4H2) was coexpressed with six different EXTs. As reported above, this P4H gene was also coexpressed with 10 different AGPs. Many peroxidase genes were coexpressed, but the greatest amount of coexpression was exhibited by At1g05240, At3g49960, At4g26010, At5g17820, and At5g67400, which were all coexpressed with eight different EXTs. Interestingly, these same peroxidase genes

Table V. *Identification, characterization, and classification of the EXT genes in Arabidopsis*

Locus Identifier[a]	Name[b]	Class	SP$_3$/SP$_4$/SP$_5$/ YXY Repeats	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top Five BLAST Hit HRGPs[e]
At1g26240	EXT20	SP$_5$	2/1/40/44	478	Yes	No	Roots	0	1/3/2/0/0	EXT17, EXT21, EXT22, LRX5, EXT1/4
At1g26250	EXT21	SP$_5$	7/0/28/40	443	Yes	No	Roots	0	1/0/5/0/3	EXT1/4, EXT17, EXT20, EXT22, LRX5
At4g08370	EXT22	SP$_5$	3/1/13/18	350	Yes	No		2	1/0/0/0/0	EXT20, EXT21, EXT17, LRX1, EXT7
At4g13390	EXT18	SP$_5$/SP$_4$	0/14/8/13	429	Yes	No	Roots	0	4/0/2/0/0	EXT11, EXT12, EXT13, EXT14, EXT15
At5g19810	**EXT19**	SP$_5$/SP$_4$	0/4/13/1	249	Yes	No	Roots	0	7/0/1/0/1	EXT39, EXT35, EXT40, AGP9C
At1g23720	EXT6	SP$_4$	2/61/3/34	895	No	No	Roots	0	1/3/0/0/0	EXT1/4
At2g24980	EXT7	SP$_4$	3/37/0/21	559	Yes	No	Roots	0	0/0/1/0/1	EXT13, EXT14, EXT11, EXT12, EXT16
At2g43150	EXT8	SP$_4$	0/22/0/9	212	Yes	No	Roots	0	1/0/1/0/1	EXT10, EXT2, EXT7, EXT6, EXT9
At3g28550	EXT9	SP$_4$	3/70/0/35	1,018	Yes	No	Roots	0	0/0/1/0/0	EXT10, EXT2, EXT6, EXT15, EXT14
At3g54580	EXT10	SP$_4$	2/68/0/33	951	Yes	No	Roots	0	0/0/1/0/2	HAE3, EXT2, EXT9, EXT1/4, PRP3
At3g54590	EXT2	SP$_4$	2/51/0/24	743	Yes	No	Roots	0	2/0/0/0/0	EXT10, EXT9, EXT1/4
At4g08400	EXT11	SP$_4$	2/31/0/26	513	Yes	No	Pollen, roots	0	2/1/1/0/0	EXT12, EXT14, EXT13, EXT16, EXT7
At4g08410	EXT12	SP$_4$	2/41/0/26	707	No	No	Roots	0	3/0/0/0/0	EXT11, EXT14, EXT13, EXT16, EXT7
At5g06630	EXT13	SP$_4$	1/29/0/17	440	Yes	No	Roots	0	3/2/1/0/0	EXT7, EXT14, EXT11, EXT12, EXT16
At5g06640	EXT14	SP$_4$	2/42/0/25	689	No	No	Roots	0	1/0/2/0/0	EXT12, EXT11, EXT13, EXT7, EXT16
At5g35190	EXT15	SP$_4$	2/12/2/8	328	Yes	No	Roots	0	2/0/1/0/1	EXT11, EXT12, EXT7, EXT13, EXT16
At5g49080	EXT16	SP$_4$	0/41/0/23	609	Yes	No	Roots	0	1/0/0/0/0	EXT11, EXT12, EXT7, EXT14, EXT13
At1g21310	EXT3/5	SP$_4$/SP$_3$	13/27/1/14	431	Yes	No	Radicle, roots	1	0/1/0/2/0	EXT1/4, HAE3
At1g76930	EXT1/4	SP$_4$/SP$_3$	8/9/0/1	293	Yes	No	Roots	3	0/1/1/1/0	EXT3/5, PAG10, PEX1, HAE3
At4g08380	EXT17	SP$_3$	34/2/0/49	437	Yes	No	Roots	0	5/1/0/0/0	EXT20, EXT22, EXT21
At1g02405	**EXT30**	Short	0/3/0/0	134	Yes	Yes	Siliques	0	1/1/1/0/2	EXT33, EXT31
At1g23040	**EXT31**	Short	0/2/0/0	144	Yes	Yes		0	4/5/1/0/1	EXT33, EXT30
At1g54215	**EXT32**	Short	0/1/1/0	169	Yes	No		0	1/1/0/0/0	LRX6, LRX3, LRX2, PRP17, PEX4
At1g70990	**EXT33**	Short	0/2/0/1	176	No	Yes	Roots	0	7/4/6/0/3	EXT31, EXT30
At3g06750	**EXT34**	Short	0/1/1/1	147	Yes	Yes		0	5/0/2/0/8	EXT41, EXT37
At3g20850	**EXT35**	Short	1/0/1/2	134	Yes	No	Roots	0	0/0/8/0/1	EXT40, EXT39
At3g49270	**EXT36**	Short	0/2/0/0	148	Yes	No	Siliques	2	1/0/2/0/55	LRX1, LRX2, EXT32, EXT19, EXT39
At4g16140	**EXT37**	Short	0/1/1/4	164	Yes	Yes		0	1/0/2/0/1	EXT41, EXT34
At5g11990	**EXT38**	Short	4/0/1/1	181	Yes	Yes		0	1/2/2/0/1	PEX4, EXT21, LRX1, LRX3, LRX5
At5g19800	**EXT39**	Short	0/0/3/1	96	Yes	No	Roots	0	1/1/0/0/0	EXT19, EXT35, EXT40
At5g26080	**EXT40**	Short	2/1/3/0	141	Yes	No	Roots	0	2/1/0/0/0	EXT35, EXT39, EXT19, PERK13, PAG10
At5g49280	**EXT41**	Short	0/2/0/2	162	Yes	Yes	nr[f]	0	2/0/1/0/1	EXT34, EXT37
At1g12040	LRX1	Chimeric	1/17/7/9	744	Yes	No	Roots	0	1/1/2/0/5	LRX2, LRX3, LRX5, LRX4, LRX7
At1g49490	PEX2	Chimeric	1/13/1/0	847	Yes	No	Pollen	0	1/0/8/0/1	PEX1, PEX3, PEX4, LRX5, LRX3
At1g62440	LRX2	Chimeric	4/12/6/3	826	No	No	Roots	2	3/1/7/5/1	LRX1, LRX5, LRX4, LRX3, LRX6

(*Table continues on following page.*)

Table V. (*Continued from previous page.*)

Locus Identifier[a]	Name[b]	Class	$SP_3/SP_4/SP_5/$ YXY Repeats	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top Five BLAST Hit HRGPs[e]
At2g15880	PEX3	Chimeric	2/16/9/1	727	No	No	Pollen	1	2/1/8/0/2	PEX4, PEX1, PEX2, LRX5, LRX3
At3g19020	PEX1	Chimeric	1/19/5/0	956	Yes	No	Pollen	0	0/3/7/0/0	PEX2, PEX3, PEX4, LRX5, LRX4
At3g22800	LRX6	Chimeric	1/0/2/6	470	Yes	No	Root	0	3/2/5/0/1	LRX3, LRX4, LRX5, LRX2, LRX1
At3g24480	LRX4	Chimeric	2/1/3/1	494	Yes	No		0	1/2/0/0/1	LRX3, LRX5, LRX2, LRX6, LRX1
At4g13340	LRX3	Chimeric	4/13/15/3	760	Yes	No		0	1/3/2/0/1	LRX4, LRX5, LRX2, LRX6, LRX1
At4g18670	LRX5	Chimeric	3/1/5/3	839	Yes	No		1	2/0/7/0/7	LRX4, LRX3, LRX2, LRX6, LRX1
At4g33970	PEX4	Chimeric	4/10/4/1	699	Yes	No	Pollen	0	2/3/7/0/1	PEX3, PEX1, PEX2, LRX4, LRX5
At5g25550	LRX7	Chimeric	1/0/1/1	433	Yes	No	Stamen	0	3/0/1/0/2	LRX4, LRX3, LRX5, LRX2, LRX1
At1g10620	PERK11	Chimeric	2/0/0/0	718	No	No	Pollen	7	9/0/11/2/1	PERK12, PERK13, PERK8, PERK6, PERK10
At1g23540	PERK12	Chimeric	1/2/0/0	720	No	No	Pollen	7	9/2/3/0/0	PERK13, PERK11, PERK8, PERK1, PERK10
At1g26150	PERK10	Chimeric	4/2/1/1	760	No	No		7	4/1/2/1/0	PERK8, PERK13, PERK12, PERK1, PERK11
At1g49270	PERK7	Chimeric	1/4/1/0	699	No	No	Pollen	6	2/1/3/3/0	PERK6, PERK5, PERK1, PERK4, PERK13
At1g52290	PERK15	Chimeric	0/0/0/0	509	No	No		7	0/5/5/2/0	PERK1, PERK5, PERK4, PERK6, PERK7
At1g70460	PERK13	Chimeric	3/2/2/0	710	No	No	Roots	7	6/4/2/1/2	PERK12, PERK11, PERK8, PERK1, PERK10
At2g18470	PERK4	Chimeric	1/0/1/1	633	No	No	Pollen	7	3/2/6/0/2	PERK1, PERK6, PERK5, PERK7, PERK3
At3g18810	PERK6	Chimeric	1/1/2/0	700	No	No	Pollen	6	9/7/2/2/1	PERK7, PERK5, PERK4, PERK1, PERK12
At3g24540	PERK3	Chimeric	0/1/1/0	509	No	No		8	0/0/5/0/0	PERK1, PERK4, PERK5, PERK6, PERK7
At3g24550	PERK1	Chimeric	3/0/0/0	652	No	No		7	5/3/2/0/0	PERK4, PERK3, PERK5, PERK6, PERK7
At4g32710	PERK14	Chimeric	0/0/0/0	388	No	No		7	0/2/4/0/2	PERK1, PERK5, PERK15, PERK7, PERK6
At4g34440	PERK5	Chimeric	2/0/0/0	670	No	No	Pollen	8	2/1/5/0/0	PERK6, PERK7, PERK1, PERK4, PERK13
At5g38560	PERK8	Chimeric	5/2/2/3	681	No	No		7	4/0/5/1/0	PERK10, PERK13, PERK12, PERK11, PERK1
At3g11030	**EXT50**	Chimeric	0/5/0/0	451	Yes	No		4	23/0/2/1/0	LRX6, LRX3, PEX2, PEX4, LRX2
At3g19430	**EXT51**	Chimeric	0/7/0/0	559	No	No	Root	2	0/0/3/0/0	LRX3, PEX3, PRP16, PEX1, LRX5
At3g53330	**EXT52**	Chimeric	0/3/0/2	310	Yes	No	nr	1	4/4/5/0/1	PAG17, PAG9, PAG10
At1g62760	**HAE1**	AGP/EXT hybrid	2/0/2/0	312	Yes	No	Pollen	0	1/3/1/0/0	LRX5, AGP54, PAG10, EXT51, AGP9
At3g50580	**HAE2**	AGP/EXT hybrid	1/2/1/0	265	Yes	No	Stamen	1	0/13/0/0/0	PRP8
At4g11430	**HAE3**	AGP/EXT hybrid	2/0/2/0	219	No	No		1	0/0/0/1/0	EXT37, LRX5, EXT19, LRX3, EXT1/4
At4g22470	**HAE4**	AGP/EXT hybrid	2/1/0/0	375	Yes	No	Leaves	0	0/2/1/0/0	PRP14, PRP16, PRP17, PRP15

[a] Italics indicates a protein that did not meet our search criteria but was identified previously in the primary literature. [b] Boldface indicates a protein that was not previously identified in the primary literature or by Johnson et al. (2003b). [c] Signal peptide. [d] Indicates the number of mutants available in each location: P, promoter; 5, 5′ UTR; E, exon; I, intron; 3, 3′ UTR. [e] Underline indicates the result of a BLAST search with filtering turned off. [f] Not reported. This indicates that data for a particular protein are not found in Genevestigator, Arabidopsis Membrane Protein Library, or MPSS.

Table VI. HRGPs, GTs, P4Hs, and peroxidases coexpressed with EXTs

HRGP Locus Identifier	Name	No. of Coexpressed EXTs	GT Locus Identifier	Name	Family	No. of Coexpressed EXTs	GT Locus Identifier Continued	Name	Family	No. of Coexpressed EXTs	P4H Locus Identifier	Name	No. of Coexpressed EXTs	Peroxidase Locus Identifier	Name	No. of Coexpressed EXTs
At4g12730	FLA2	15	At2g32620	CslB04	GT2	9	At4g36890		GT43	2	At3g06300	P4H2	6	At1g05240		8
At1g03870	FLA9	14	At1g24170		GT8	7	At4g38040		GT47	2	At2g17720	P4H5	2	At3g49960		8
At3g28550	EXT9	11	At1g74380		GT34	7	At5g03760	CslA09	GT2	2	At2g43080	P4H1	1	At4g26010	ATP13a	8
At5g06630	EXT13	11	At4g15290	CslB05	GT2	7	At5g05860		GT1	2	At5g18900	P4H4	1	At5g17820	PER57	8
At5g06640	EXT14	11	At5g22940		GT47	7	At5g07720		GT34	2				At5g67400	PER73	8
At1g23720	EXT6	10	At3g18170		GT61	6	At5g14850		GT22	2				At1g30870		7
At3g24480	LRX4	10	At3g24040		GT14	6	At5g15050		GT14	2				At3g28200		7
At3g54580	EXT10	10	At4g39350	CesA02	GT2	6	At5g16910	CslD2	GT2	2				At5g22410		6
At3g54590	EXT2	10	At1g08660		GT29	5	At5g20830		GT4	2				At4g33420		4
At4g08410	EXT12	9	At1g13250		GT8	5	At5g24300		GT5	2				At5g39580		4
At4g26320	AGP13	9	At3g61130		GT8	5	At5g41460		GT31	2				At1g77490	tAPX	3
At2g24980	EXT7	8	At4g00300		GT31	5	At5g44820		GT77	2				At2g25080	GPX1	3
At3g19430	EXT51	8	At4g01750		GT77	5	At5g61840		GT47	2				At4g09010		3
At5g10430	AGP4C	8	At5g05170	CesA03	GT2	5	At1g03520		GT14	1				At4g37530		3
At3g62680	PRP3	7	At1g02730	CslD5	GT2	4	At1g05570	Gsl06	GT48	1				At5g19890		3
At1g26250	EXT21	6	At1g27120		GT31	4	At1g06780		GT8	1				At5g40150		3
At2g43150	EXT8	6	At2g03220	FUT1	GT37	4	At1g16900		GT22	1				At1g05260	PER3	2
At2g45470	FLA8	6	At2g31790		GT1	4	At1g19360		GT77	1				At2g31570		2
At3g11700	FLA18	5	At3g03050	CslD3	GT2	4	At1g19710		GT4	1				At3g03670		2
At3g13520	AGP12P	5	At3g05320		GT65	4	At1g23870		GT20	1				At3g63080		2
At4g16980	AGP58C	5	At3g28180	CslC04	GT2	4	At1g27440		GT47	1				At2g18140		1
At5g44130	FLA13	5	At4g38240		GT13	4	At1g32900		GT5	1				At2g22420	PER17	1
At5g53250	AGP22P	5	At5g05890		GT1	4	At1g34130		GT66	1				At2g37130	PER21	1
At1g52290	PERK15	4	At5g09870	CesA05	GT2	4	At1g34270		GT47	1				At2g41480		1
At1g55330	AGP21P	4	At5g47780		GT8	4	At1g50580		GT1	1				At2g43480		1
At2g10940	PRP15	4	At5g64740	CesA06	GT2	4	At1g51210		GT1	1				At3g21770	PER30	1
At3g06750	EXT34	4	At1g18580		GT8	3	At1g68020		GT20	1				At4g08770		1
At4g13340	LRX3	4	At1g23480	CslA03	GT2	3	At1g70090		GT8	1				At4g11290		1
At4g27520	PAG10	4	At1g24070	CslA10	GT2	3	At1g71220		GT24	1				At4g11600		1
At4g37450	AGP18K	4	At1g70290		GT20	3	At1g74800		GT31	1				At4g35000	APX3	1
At1g21310	EXT3/5	3	At1g73160		GT4	3	At2g25300		GT31	1				At5g24070		1
At1g28290	AGP31I	3	At2g22900		GT34	3	At2g31960	Gsl03	GT48	1				At5g64120		1
At2g04780	FLA7	3	At2g30150		GT1	3	At2g32530	CslB02	GT2	1						
At3g45230	AGP57C	3	At3g62720		GT34	3	At2g35610		GT77	1						
At4g16140	EXT37	3	At4g02130		GT8	3	At3g01180		GT5	1						
At4g31840	PAG13	3	At4g07960	CslC12	GT2	3	At3g07020		GT1	1						
At4g32710	PERK14	3	At5g19690		GT66	3	At3g10630		GT4	1						
At5g11740	AGP15P	3	At5g39990		GT14	3	At3g15350		GT14	1						
At5g21160	AGP32I	3	At5g65685		GT5	3	At3g21750		GT1	1						
At5g55730	FLA1	3	At5g66690		GT1	3	At3g28340		GT8	1						
At5g56540	AGP14P	3	At1g06000		GT1	2	At3g45100		GT4	1						
At2g21140	PRP2	2	At1g07240		GT1	2	At3g55710		GT1	1						
At2g33790	AGP30I	2	At1g10400		GT1	2	At3g59100	Gsl11	GT48	1						
At2g47930	AGP26C	2	At1g30530		GT1	2	At4g01220		GT77	1						

(*Table continues on following page.*)

Table VI. (*Continued from previous page.*)

HRGP Locus Identifier	Name	No. of Coexpressed EXTs	GT Locus Identifier	Name	Family	No. of Coexpressed EXTs	GT Locus Identifier Continued	Name	Family	No. of Coexpressed EXTs	P4H Locus Identifier	Name	No. of Coexpressed EXTs	Peroxidase Locus Identifier	Name	No. of Coexpressed EXTs
At3g52370	FLA15	2	At1g43620		GT1	2	At4g04970	Gsl01	GT48	1						
At3g61640	AGP20P	2	At1g53290		GT31	2	At4g09500		GT1	1						
At4g09030	AGP10C	2	At1g71070		GT14	2	At4g15550		GT1	1						
At4g18670	LRX5	2	At1g78580		GT20	2	At4g16600		GT8	1						
At5g15780	PRP11	2	At2g15370	FUT5	GT37	2	At4g18230		GT1	1						
At5g40730	AGP24P	2	At2g18700		GT20	2	At4g21060		GT31	1						
At1g26150	PERK10	1	At2g20370		GT47	2	At4g23490		GT31	1						
At1g26240	EXT20	1	At2g20810		GT8	2	At4g24000	CslG2	GT2	1						
At1g62500	PRP14	1	At2g24630	CslC08	GT2	2	At4g31590	CslC05	GT2	1						
At1g70990	EXT33	1	At2g35650	CslA07	GT2	2	At5g01220		GT4	1						
At2g22470	AGP2C	1	At2g37585		GT14	2	At5g11110		GT4	1						
At2g25060	PAG2	1	At2g44660		GT57	2	At5g12890		GT1	1						
At2g35860	FLA16	1	At3g02350		GT8	2	At5g15650		GT75	1						
At3g22120	PRP16	1	At3g16520		GT1	2	At5g22740	CslA02	GT2	1						
At3g22800	LRX6	1	At3g27540		GT17	2	At5g38460		GT57	1						
At3g24550	PERK1	1	At3g29320		GT35	2	At5g50420		GT68	1						
At3g60900	FLA10	1	At3g46970		GT35	2	At5g62220		GT47	1						
At5g14920	PRP18	1	At3g56000	CslA14	GT2	2										
At5g25090	PAG15	1	At4g02500		GT34	2										
At5g53870	PAG17	1	At4g18240		GT5	2										
At5g60490	FLA12	1	At4g31780		GT28	2										
At5g64310	AGP1C	1	At4g32120		GT31	2										
At5g65390	AGP7C	1	At4g32410	CesA01	GT2	2										

Table VII. *PRPs identified from the Arabidopsis genome based on biased amino acid composition and repeat units*

The number in parentheses indicates the number of proteins that had a predicted signal peptide sequence.

Search Criteria	Total	PRPs	AGP	EXT	Hybrid	Other
≥45% PVKCYT	113 (64)	15 (14)	10 (10)	31 (26)	3 (2)	54 (12)
Two or more KKPCPP	2 (2)	2 (2)	0	0	0	0
Two or more PPVX[KT]	13 (11)	7 (7)	2 (2)	1 (1)	1 (0)	2 (1)

were coexpressed with the greatest number of AGP genes as well (Table III). Given that EXTs are known to be cross-linked at YXY sequence motifs by an EXT peroxidase with an acidic pI, it was interesting to observe that the At3g03670-encoded peroxidase, which had a predicted endomembrane localization and a predicted pI of 4.8, was coexpressed with two of the three EXTs containing the greatest numbers of YXY sequence repeats (i.e. EXT20 and EXT21).

EXT Gene Organization and Mutants

Information was extracted from the TAIR and SALK Web sites with regard to the gene structure and available genetic mutants for each of the predicted EXTs. With the exception of the PERK genes, EXT genes including the four HRGP hybrid genes contain few, if any, introns (Table V; Supplemental Table S6). Of the 46 non-PERK EXT genes, 36 had no introns and eight had only one or two introns. All four HAEs contained either zero or one intron. One chimeric EXT (At3g11030), however, was predicted to have four introns. In contrast, the PERK genes contained between six and eight introns.

Examination of the various mutant lines available for research showed that all of the EXT genes (including HAEs) had one or more mutants available. Of these mutants, 29% are in the promoter region, 17% are in the 5′ UTR, 30% are in an exon, 4% are in an intron, and 20% are in the 3′ UTR (Table V; Supplemental Table S7).

Finding and Classifying PRPs

The BIO OHIO program was used to identify potential PRPs primarily by searching for proteins with a biased amino acid composition of at least 45% PVKCYT. In addition, PRPs were identified by searching for KKPCPP and PPVX(K/T) sequences repeated two or more times (Fowler et al., 1999). The program initially identified 113 potential PRPs by searching for 45% PVKCYT and identified 13 and two potential PRPs by searching for the PPVX(K/T) and KKPCPP repeats, respectively. Eleven of these 13 potential PRPs and both of these two potential PRPs were also identified with the 45% PVKCYT search criteria (Table VII).

The 113 proteins identified by the program were further examined individually to determine if they appeared to be PRPs. The presence of a signal peptide was one such factor, as was the presence and location of PPV repeats, since these peptide sequences are often present in known PRPs. The PRPs, like the EXTs, are not known to contain GPI anchor addition sequences, but the presence of such sequences was queried nonetheless. By these criteria, 15 of the 113 were classified as PRPs. The 45% PVKCYT search criteria failed to find all the potential PRP sequences and had a high rate of false positives. In addition to the 15 PRPs, nine AGPs (AGP45P, AGP56C, AGP9C, AGP7C, AGP4C, AGP18K, AGP19K, AGP30I, AGP33I), 31 EXTs (EXT40, EXT17, EXT32, EXT37, EXT41, LRX3, LRX1, EXT39, EXT20, EXT21, EXT3/5, EXT8, EXT7, EXT35, EXT9, EXT10, EXT2, EXT11, EXT13, EXT16, EXT15, EXT18, EXT1/4, EXT22, EXT19, EXT30, PEX3, EXT6, EXT12, EXT14, EXT51), and three hybrid HRGPs (HAE2, HAE3, HAE4) were found with the 45% PVKCYT search. In addition, two AGPs (AGP4C, AGP9C), one EXT (EXT1/4), and one hybrid HRGP (HAE3) were found with the two PPVX(K/T) repeat search; further information on these sequences was presented in the AGP and EXT sections above. Three additional PRPs (PRP8, PRP9, PRP11) did not pass the biased amino acid test but were found instead by a database annotation search. The locus identifiers of these sequences are indicated in italics in Table VIII. With these additional PRPs, 18 total PRPs were found and subjected to further analysis. Six of the 18 PRPs contained a non-HRGP domain along with a PRP domain and thus were classified as chimeric PRPs. The remaining 12 PRPs were not divided further into subclasses (Table VIII). Representative sequences of these two classes of PRPs are shown in Figure 4.

BLAST analysis was conducted to identify other potential PRP sequences and to provide insight to PRP sequences with the greatest similarity (Table VIII; Supplemental Table S1). BLAST was somewhat successful in identifying other PRPs, but all PRPs cannot be found with a single BLAST search. Interestingly, the BLAST searches showed that six of the 18 PRPs are similar to AGP30, a nonclassical (chimeric) AGP. In fact, when AGP30 was used as the query sequence in a BLAST search, the top four hits were all PRPs rather than AGPs (Table II; Supplemental Table S1). Also consistent with these findings is the fact that AGP30 was not identified with the traditional 50% PAST search used for AGPs but was found with the 45% PVKCYT search used for PRPs.

The PRPs ranged in size from 126 to 761 amino acids (Table VIII; Supplemental Fig. S11). Eleven of the 12 PRPs were predicted to have a signal peptide, but

Table VIII. *Identification, characterization, and classification of the PRP genes in Arabidopsis*

Locus Identifier[a]	Name[b]	Class	PPVX[KT]/ KKPCPP/ PPV Repeats	Amino Acids	SP[c]	GPI	Organ-Specific Expression	Introns	P/5/E/I/3 Mutants[d]	Top Five BLAST Hits HRGPs[e]
At1g15825	**PRP5**	PRP	1/0/4	126	No	No		0	0/1/3/0/8	AGP9C, PRP6, PRP11, PRP4, AGP19K
At1g54970	PRP1	PRP	13/0/2	335	Yes	No	Roots	1	0/0/1/1/0	PRP3, PRP7, AGP31I, PRP6, PRP16
At2g21140	PRP2	PRP	0/4/7	321	Yes	No		1	6/1/0/0/0	PRP4
At2g27380	**PRP6**	PRP	22/0/24	761	Yes	No	Endosperm	0	6/1/1/0/2	EXT6, EXT10, EXT9, PEX1, EXT2
At2g47530	**PRP7**	PRP	0/0/0	184	Yes	No	Roots	1	3/6/5/0/1	PRP1, PRP3, AGP31I, AGP30I
At3g50570	**PRP8**	PRP	0/0/0	189	Yes	No	Stamen	0	3/0/3/0/2	HAE2
At3g62680	PRP3	PRP	14/0/0	313	Yes	No	Roots	1	2/3/7/0/5	PRP1, PRP7, AGP30I, AGP31I, EXT1/4
At4g38770	PRP4	PRP	0/7/14	448	Yes	No		1	7/0/5/1/3	PRP2, AGP55C
At5g09520	**PRP9**	PRP	0/0/0	130	Yes	No	Radicle, root	0	5/2/1/0/1	PRP10, AGP1C
At5g09530	**PRP10**	PRP	0/0/0	360	Yes	No	Radicle, root	0	3/0/7/0/7	PRP9, PRP11, PRP4, PRP15, PRP16
At5g15780	**PRP11**	PRP	0/0/3	401	Yes	No		1	1/2/6/2/2	AGP31I, AGP30I, PRP1
At5g59170	**PRP12**	PRP	0/0/5	288	Yes	No	Seeds	0	5/4/5/0/0	AGP55C
At1g09460	**PRP13**	Chimeric	2/0/4	330	Yes	Yes		2	3/0/1/0/0	EXT51, AGP9C, PRP18, PRP16, PERK8
At1g62500	**PRP14**	Chimeric	4/0/4	297	Yes	No	Shoot apex	0	6/1/0/0/0	PRP10, PRP9, PRP11, PRP4, PRP15
At2g10940	**PRP15**	Chimeric	0/0/11	291	Yes	No		1	2/2/1/0/2	PRP14, PRP16, PRP17, HAE4, AGP2C
At3g22120	**PRP16**	Chimeric	7/0/0	334	Yes	No		0	2/0/1/0/0	PRP17, PRP14, PRP15, HAE4
At4g15160	**PRP17**	Chimeric	1/0/0	428	Yes	No		3	2/1/4/3/2	PRP16, PRP14, PRP15
At5g14920	**PRP18**	Chimeric	2/0/7	275	Yes	No	Petiole	3	2/1/4/1/1	PRP6, AGP31I, PRP16, EXT51, PEX3

[a] Italics indicates a protein found using the Arabidopsis database annotation search. [b] Boldface indicates a protein that was not previously identified in the primary literature. [c] Signal peptide. [d] Indicates the number of mutants available in each location: P, promoter; 5, 5′ UTR; E, exon; I, intron; 3, 3′ UTR. [e] Underline indicates the result of a BLAST search with filtering turned off.

none was predicted to have a GPI anchor. The six chimeric PRPs ranged in size from 275 to 428 amino acids. All six chimeric PRPs were predicted to have a signal peptide, and one was predicted to have a GPI anchor.

PRP Gene Expression and Coexpressed HRGPs, GTs, P4Hs, and Peroxidases

In order to elucidate patterns of gene expression for these predicted PRPs, the same three public databases were searched as with the AGPs and EXTs. While most PRPs had a broad range of expression throughout the plant, several of the PRP genes showed organ-specific expression. Notably, several PRPs were specifically or preferentially expressed in the roots, while other individual PRPs were expressed in the endosperm, shoot apex, and petiole (Table VIII; Supplemental Figs. S12–S15). Moreover, in examining the expression levels of all the PRP genes, endosperm-specific At2g27380 (PRP6) was the most highly expressed one, as indicated by its high relative signal intensity.

Unlike the AGPs and EXTs, the PRPs displayed some common and dramatic (i.e. approximately 8-fold or more) patterns of environmental stress-induced gene expression. For example, eight of the PRP genes (PRP1, -2, -8, -3, -4, -9, -10, and -15) were down-regulated by ABA, while two of the PRP genes (PRP6 and -14) were up-regulated by ABA. In addition, three PRPs (PRP2, -3, and -11) were up-regulated by zeatin, three PRPs (PRP 4, -11, and -16) were up-regulated by nematode infection, and two PRPs (PRP9 and -10) were up-regulated by *Pseudomonas syringae* infection.

Next, the PRP genes were examined with respect to coexpressed genes using The Arabidopsis Co-Response Database (Table IX; Supplemental Table S8). Twelve out of the 18 PRPs had data available. In analyzing the data, a focus was placed not only on other HRGPs but on GTs, P4Hs, and peroxidases, since these enzymes are responsible for posttranslational modification of PRPs; this approach represents one potential avenue to identify genes involved in the posttranslational modification of PRPs. In terms of PRPs being expressed with other HRGPs, 46 different HRGPs are coexpressed with at least one PRP. The HRGP showing greatest coexpression was FLA8, which was coexpressed with five PRPs; FLA8 was

PRPs

```
>At2g27380-PRP6
MRVPLIDFLRFLVLILSLSGASVAADATVKQNFNKYETDSGHAHPPPIYGAPPSYTTPPPPIYSPPIYPPPIQKPPTYSPPIYPPPIQK
PPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPTYSPPIYPPPIQKPPTPSYSPPVKPPPVQMPPTPTYSPPIKPPPVHKPPTP
TYSPPIKPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPPIKPPPVHKPPTPTYSPPVKPPPVHKPPTPIYSPPIKPPPVHKPPTPIYSPP
VKPPPVQTPPTPIYSPPVKPPPVHKPPTPTYSPPVKSPPVQKPPTPTYSPPIKPPPVQKPPTPTYSPPIKPPPVKPPTPIYSPPVKPPP
VHKPPTPIYSPPVKPPPVHKPPTPIYSPPVKPPPIQKPPTPTYSPPIKPPPLQKPPTPTYSPPIKLPPVKPPTPIYSPPVKPPPVHKPP
TPIYSPPVKPPPVHKPPTPTYSPPIKPPPVKPPTPTYSPPVQPPPVQKPPTPTYSPPVKPPPIQKPPTPTYSPPIKPPPVKPPTPTYSP
PIKPPPVHKPPTPTYSPPIKPPPIHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKP
PPVHKPPTPTYSPPIKPPPVHKPPTPTYSPPIKPPPVQKPPTPTYSPPVKPPPVQLPPTPTYSPPVKPPPVQVPPTPTYSPPVKPPPVQ
VPPTPTYSPPIKPPPVQVPPTPTTPSPPQGGYGTPPPYAYLSHPIDIRN
```

```
>At4g38770-PRP4
MRILPEPRGSVPCLLLLVSVLLSATLSLARVVEVVGYAESKIKTPHAFSGLRVTIDCKVNKGHFVTKGSGNIDDKGKFGLNIPHDIVSD
NGALKEECYAQLHSAAGTPCPAHDGLESTKIVFLSKSGDKHILGLKQNLKFSPEICVSKFFWPMPKLPPFKGFDHPFPLPPPLELPPFL
KKPCPPKYSPPVEVPPPVPVYEPPPKKEIPPPVPVYDPPPKKEVPPPVPVYKPPPKVELPPPIPKKPCPPKPPKIEHPPPVPVYKPPPK
IEKPPPVPVYKPPPKIEHPPPVPVHKLPKKPCPPKKVDPPPVPVHKPPTKKPCPPKKVDPPPVPVHKPPPKIVIPPPKIEHPPPVPVYK
PPPKIEHPPIYIPPIVKKPCPPPVPIYKPPVVIPKKPCPPPVPVYKPPVVVIPKKPCPPLPQLPPLPKFPPLPPKYIHHPKFGKWPPLP
PHP
```

Chimeric PRP

```
>At5g14920-PRP18
MALSLLSVFIFFHVFTNVVFAASNEESNALVSLPTPTLPSPSPATKPPSPALKPPTPSYKPPTLPTTPIKPPTTKPPVKPPTIPVTPVK
PPVSTPPIKLPPVQPPTYKPPTPTVKPPSVQPPTYKPPTPTVKPPTTSPVKPPTTPPVQSPPVQPPTYKPPTSPVKPPTTTPPVKPPTT
TPPVQPPTYNPPTTPVKPPTAPPVKPPTPPPVRTRIDCVPLCGTRCGQHSRKNVCMRACVTCCYRCKCVPPGTYGNKEKCGSCYANMKT
RGGKSKCP
```

Figure 4. Protein sequences encoded by representative PRP gene classes in Arabidopsis. Colored sequences at the N terminus indicate predicted signal peptide (green). PPVX(K/T) (gray), KKPCPP (teal), and PPV (pink) repeats are also indicated.

also coexpressed with 16 AGPs. FLA9 and FLA2, which were coexpressed with many AGPs and EXTs, were each coexpressed with three PRPs. For the GTs, At5g22940 of the GT47 family was coexpressed with six PRPs, twice as many as any other GT. Moreover, At1g24170, a GT8 family member that was coexpressed with many AGPs and EXTs, was not coexpressed with any PRPs. At3g14570 (Gsl04), a member of the GT family 48, was coexpressed with three PRPs; it was also coexpressed with four AGPs but no EXTs. For the P4Hs, two of 13 members of the P4H gene family, At3g06300 (P4H2) and At5g18900 (P4H4), were coexpressed with two and one PRPs, respectively, as well as with many AGPs and EXTs. For the peroxidases, some peroxidase genes were coexpressed. The greatest amount of coexpression was exhibited by At1g77490 (tAPX) and At2g22420 (PER17); each was coexpressed with two PRPs. Both of these peroxidases also were coexpressed with EXTs and AGPs.

PRP Gene Organization and Mutants

Information was extracted from the TAIR and SALK Web sites with regard to the gene structure and available genetic mutants for each of the predicted PRP genes. None of the 18 PRPs contained more than three introns, with most containing either zero (eight of 18) or one intron (seven of 18; Table VIII; Supplemental Table S9).

Examination of the various mutant lines available for research showed that all of the PRP genes have one or more mutants available. Of these mutants, 32% were in the promoter region, 14% were in the 5′ UTR, 30% were in an exon, 4% were in an intron, and 20% were in the 3′ UTR (Table VIII; Supplemental Table S10).

DISCUSSION

The BIO OHIO Program for Finding and Analyzing HRGP Genes Based on Biased Amino Acid Compositions and Amino Acid Sequence Motifs

As genomes are sequenced, bioinformatic tools need to be developed to analyze such data efficiently and accurately. Here, we describe one such tool for the purpose of identifying and analyzing HRGPs encoded by nucleic acid sequences. The BIO OHIO software has the ability to identify AGPs, EXTs, and PRPs as well as hybrid and chimeric HRGPs. This program requires only that the protein sequence data be available as a data file, which is routinely generated in a completed genome sequencing project. Here, the BIO OHIO program was used to search the 28,952 protein sequences encoded by the Arabidopsis genome. Several different strategies were used by the program to identify candidate HRGPs. Specifically, the program has the ability to identify proteins meeting a user-defined amino acid

Table IX. *HRGPs, GTs, P4Hs, and peroxidases coexpressed with PRPs*

HRGP Locus Identifier	Name	No. of Coexpressed PRPs	GT Locus Identifier	Name	Family	No. of Coexpressed PRPs	P4H Locus Identifier	Name	No. of Coexpressed PRPs	Peroxidase Locus Identifier	Name	No. of Coexpressed PRPs
At2g45470	FLA8	5	At5g22940		GT47	6	At3g06300	P4H2	2	At1g68850		2
At4g16980	AGP58C	4	At3g14570	Gsl04	GT48	3	At5g18900	P4H4	1	At1g77490	tAPX	2
At1g03870	FLA9	3	At1g07250		GT1	2				At2g22420	PER17	2
At1g52290	PERK15	3	At1g08660		GT29	2				At1g05240		1
At2g47930	AGP26C	3	At3g29320		GT35	2				At1g30870		1
At4g12730	FLA2	3	At3g46970		GT35	2				At1g71695	PER12	1
At2g04780	FLA7	2	At4g02500		GT34	2				At2g25080	GPX1	1
At3g06750	EXT34	2	At4g31780		GT28	2				At2g31570		1
At4g18670	LRX5	2	At4g39350	CesA02	GT2	2				At3g21770	PER30	1
At4g26320	AGP13P	2	At5g03760	CslA09	GT2	2				At3g28200		1
At4g37450	AGP18K	2	At5g05890		GT1	2				At3g49120		1
At5g55730	FLA1	2	At5g22740	CslA02	GT2	2				At3g49960		1
At5g56540	AGP14P	2	At5g50420		GT68	2				At4g08770		1
At1g09460	PRP13	1	At1g06780		GT8	1				At4g09010		1
At1g23720	EXT6	1	At1g11720		GT5	1				At4g26010	ATP13a	1
At1g26150	PERK10	1	At1g13250		GT8	1				At5g17820	PER57	1
At1g28290	AGP31I	1	At1g16570		GT33	1				At5g22410		1
At2g10940	PRP15	1	At1g19360		GT77	1				At5g67400	PER73	1
At2g24980	EXT7	1	At1g21480		GT47	1						
At2g33790	AGP30I	1	At1g23480	CslA03	GT2	1						
At2g35860	FLA16	1	At1g27440		GT47	1						
At3g11700	FLA18	1	At1g71220		GT24	1						
At3g19430	EXT51	1	At1g78580		GT20	1						
At3g22120	PRP16	1	At2g03220	FUT1	GT37	1						
At3g24480	LRX4	1	At2g22900		GT34	1						
At3g52370	FLA15	1	At2g29750		GT1	1						
At3g54590	EXT2	1	At2g31790		GT1	1						
At3g60900	FLA10	1	At2g32620	CslB04	GT2	1						
At4g08410	EXT10	1	At2g35650	CslA07	GT2	1						
At4g09030	AGP10C	1	At3g06440		GT31	1						
At4g13340	LRX3	1	At3g18170		GT61	1						
At4g15160	PRP17	1	At3g24040		GT14	1						
At4g16140	EXT37	1	At3g45100		GT4	1						
At4g27520	PAG10	1	At3g59100	Gsl11	GT48	1						
At5g06630	EXT13	1	At3g61130		GT8	1						
At5g06640	EXT14	1	At4g02130		GT8	1						
At5g09520	PRP9	1	At4g07960	CslC12	GT2	1						
At5g09530	PRP10	1	At4g15290	CslB05	GT2	1						
At5g10430	AGP4C	1	At4g18240		GT5	1						
At5g14920	PRP18	1	At4g38040		GT47	1						
At5g15780	PRP11	1	At4g38270		GT8	1						
At5g18690	AGP25C	1	At5g05170	CesA03	GT2	1						
At5g21160	AGP32I	1	At5g15650		GT75	1						
At5g40730	AGP24P	1	At5g16190	CslA11	GT2	1						
At5g53250	AGP22P	1	At5g16510		GT75	1						
At5g60490	FLA12	1	At5g17420	CesA07	GT2	1						
			At5g19690		GT66	1						
			At5g24300		GT5	1						
			At5g41460		GT31	1						
			At5g47780		GT8	1						
			At5g53340		GT31	1						
			At5g54690		GT8	1						

composition in full-length proteins or proteins of some defined size. This strategy was effective in identifying candidate classical AGPs, Lys-rich AGPs, AG peptides, and certain PRPs. The program can also be used to identify proteins containing specific, user-defined peptide sequences repeated any number of times. This strategy was used to identify candidate FLAs, EXTs, and certain PRPs. Both strategies were able to identify candidate hybrid and chimeric HRGPs. Another search strategy built into the program is to search for keywords within the annotated Arabidopsis protein database. This approach proved useful in finding

Table X. *A summary of the HRGP superfamily in Arabidopsis*

Boldface entries are subtotals for the various HRGP families.

HRGP Family	HRGP Subfamily	Predicted No. of:		
		Genes	Signal Peptides	GPI Anchors
AGPs	Classical AGPs	22	19	14
AGPs	Lys-rich classical AGPs	3	3	2
AGPs	AG peptides	16	16	12
AGPs	(Chimeric) FLAs	21	20	10
AGPs	(Chimeric) PAGs	17	17	16
AGPs	Other chimeric AGPs	6	5	1
AGPs	**All AGP subfamilies**	**85**	**80**	**55**
EXTs	SP_5 EXTs	3	3	0
EXTs	SP_5/SP_4 EXTs	2	2	0
EXTs	SP_4 EXTs	12	9	0
EXTs	SP_4/SP_3 EXTs	2	2	0
EXTs	SP_3 EXT	1	1	0
EXTs	Short EXTs	12	11	7
EXTs	(Chimeric) LRXs	11	9	0
EXTs	(Chimeric) PERKs	13	0	0
EXTs	Other chimeric EXTs	3	2	0
EXTs	**All EXT subfamilies**	**59**	**39**	**7**
Hybrid	HAE (AGP/EXT)	4	3	0
Hybrid	**All hybrid HRGPs**	**4**	**3**	**0**
PRPs	PRPs	12	11	0
PRPs	Chimeric PRPs	6	6	1
PRPs	**All PRP subfamilies**	**18**	**17**	**1**
Total	**All AGPs, EXTs, and PRPs**	**166**	**139**	**63**

some chimeric AGPs and PRPs not identified by the above approaches. In addition, the program can search for signal peptide sequences, GPI anchor addition sequences, and repeating sequences within proteins; such additional information in conjunction with careful examination of the protein sequence was used to manually identify candidate proteins as HRGPs. In total, this bioinformatics approach identified 166 candidate HRGPs, including 85 AGPs (22 classical AGPs, three Lys-rich AGPs, 16 AG peptides, 21 [chimeric] FLAs, 17 [chimeric] PAGs, and six other chimeric AGPs), 59 EXTs (three SP_5 EXTs, two SP_5/SP_4 EXTs, 12 SP_4 EXTs, two SP_4/SP_3 EXTs, one SP_3 EXT, 12 short EXTs, 11 [chimeric] LRXs, 13 [chimeric] PERKs, and three other chimeric EXTs),18 PRPs (12 PRPs and six chimeric PRPs), and four AGP/EXT HAEs (Table X).

This bioinformatics approach has advantages over conventional BLAST searches in terms of speed and accuracy. BLAST searches are time-consuming, requiring much postanalysis data acquisition and analysis after a list of "hits" to a query sequence is obtained. Furthermore, BLAST analyses fail to identify all members of an AGP, EXT, or PRP subfamily, since many of the subfamily members have limited amino acid sequence similarities and/or have various repeated amino acid sequence modules within a given sequence, complicating the alignment process. Nonetheless, BLAST analysis was used here to identify the most closely related sequences to a given HRGP, and by playing a version of the six degrees of separation game, it could be used to identify many, but not all, HRGP members in a time-consuming, convoluted, and laborious endeavor.

Schultz et al. (2002) previously utilized a bioinformatics approach to identify candidate AGP genes from Arabidopsis. In contrast to this study, only 52 AGPs (14 classical AGPs, three Lys-rich AGPs, 10 AG peptides, 21 [chimeric] FLAs, and four other chimeric AGPs) were identified. The additional AGPs found in this study are largely attributed to using an updated Arabidopsis protein database, altering the definition of an AG peptide to include up to 90 amino acids (compared with 75), and analyzing HRGP-related sequences based on annotations in the database. In addition, Schultz et al. (2002) also identified 19 candidate EXT genes as a by-product of searching for AGPs using the greater than 50% PAST amino acid bias. As explained by Johnson et al. (2003b), these 19 genes were subsequently examined for the presence of a signal peptide and SP_3 and SP_4 repeat units. In contrast, the additional EXTs found in this study are largely attributed to using an updated protein database, to searching for SP_3 and SP_4 repeats in all the proteins encoded by the genome (not just those proteins passing the 50% PAST test), and to analyzing HRGP-related sequences based on annotations in the database and literature. Johnson et al. (2003b) also reported the existence of 17 PRPs based on searching for proteins with greater than 49% PKVY and greater than 47% PKVL amino acid biases, similar to the findings obtained in this study.

While most of the AGP, EXT, and PRP genes fitting canonical sequencing parameters are now identified, identifying chimeric HRGPs, particularly chimeric AGPs, remains a challenge, given that no clear consensus sequence exists as for the AGPs. Thus, while we have identified six chimeric AGPs in addition to the FLAs and PAGs, it is likely that other proteins contain AGP modules. For instance, two homologous Arabidopsis genes, At5g64080 and At2g13820, designated Arabidopsis *XYLOGEN PROTEIN1* (*AtXYP1*) and *AtXYP2*, respectively, are known to contain AGP-like regions, but they were not identified in our searches. A glimpse of other such chimeric AGPs was provided in a previous study, where putative GPI-anchored proteins were identified by bioinformatics to reveal not only numerous GPI-anchored AGPs but also approximately 50 other proteins containing AGP sequence modules, but annotated as phytocyanins, stellacyanin-like, uclacyanin-like, early nodulin-like, COBRA, β-(1,3)-glucanases, aspartyl proteases, LTPL, SKU5, receptor-like kinases, and other unknown or hypothetical proteins (Borner et al., 2003).

In order to identify such chimeric AGPs, the sliding windows feature of the BIO OHIO program was utilized. Specifically, the Arabidopsis protein database was searched using windows of 10, 20, and 30 amino acids and searching for greater than 80%, 90%, and 95% PAST. In order to find all 85 AGPs identified in our searches with a sliding windows approach, an amino acid composition of greater than 60% PAST is required with a window size of 10 amino acids. While this approach finds all of the AGPs predicted by our searches, it produces many false positives in the process, making this approach of limited usefulness in initial searches on its own. However, the sliding windows feature is especially useful to identify single or multiple AGP modules in chimeric AGPs when identified by other approaches.

Laboratory experimentation has verified and validated this in silico approach to identifying HRGPs. With respect to the AGPs, reports on several cloned AGP genes and/or characterized AGP glycoproteins in Arabidopsis exist and substantiate predictions made by the program (Schultz et al., 2000, 2004; Johnson et al., 2003a; van Hengel and Roberts, 2003; Sun et al., 2005; Liu and Mehdy, 2007; Yang et al., 2007). Moreover, at the protein level, several of the AGPs predicted here to have signal peptides and GPI anchors are substantiated in these reports. With respect to the EXTs, only three nonchimeric EXT genes (EXT1/4, EXT2, EXT3/5) and several LRXs and PERKs are cloned (Merkouropoulos et al., 1999; Yoshiba et al., 2001; Baumberger et al., 2003b; Nakhamchik et al., 2004). Moreover, both the LRXs and PERKs were previously examined using BLAST and other homology-based genomic tools to identify members of these two chimeric EXT classes, in agreement with the bioinformatics findings presented here (Baumberger et al., 2003a; Nakhamchik et al., 2004). In contrast to the AGPs, there is little information on the EXTs at the glycoprotein level in Arabidopsis. With respect to the PRPs, only four PRPs are cloned in Arabidopsis, namely PRP1, -2, -3, and -4, and little is known about any of the Arabidopsis PRPs from glycoprotein studies (Fowler et al., 1999). Thus, this work extends and consolidates the experimental inventory of HRGPs and makes testable predictions with respect to the presence (or absence) of signal peptides and GPI anchor addition sequences. Although the majority of HRGPs identified by this bioinformatics approach contain signal peptides, several HRGPs do not. It is unknown whether this represents limitations to the predictive power of the program or is due to the possibility that HRGPs lacking such a sequence remain inside the cells or are secreted by an alternative secretory pathway, as reported in some cases (Nickel, 2003; Lee et al., 2004). For instance, all PERKs lack a signal peptide but are localized to the plasma membrane, with the EXT region extending into the cell wall (Nakhamchik et al., 2004). Similarly, while GPI anchors predicted for many AGPs are experimentally verified in several instances, including in Arabidopsis, it was surprising to observe here and elsewhere that several EXTs and one PRP also have predicted GPI anchor addition sequences (Borner et al., 2003), which await biochemical and functional verification at the protein and cell biology levels, respectively.

Four hybrid HRGPs containing AGP and EXT sequence motifs also are encoded by the Arabidopsis genome. These hybrids, like the chimeric HRGPs, complicate the classification system. Indeed, it is human nature to classify things into discrete categories, but the chimeric and hybrid HRGPs remind us that nature cares little for the organizational principles coveted by the human mind. Consequently, it is perhaps best to view the HRGPs as a spectrum of molecules composed of some combination of hyperglycosylated AGP modules, moderately glycosylated EXT modules, lightly glycosylated or nonglycosylated PRP modules, and, in the case of chimeric HRGPs, other non-HRGP modules.

HRGP Gene Expression in Development and in Response to Biotic and Abiotic Stress

Microarray as well as MPSS data are valuable, publicly available genetic resources for the Arabidopsis community, effectively revealing developmental, organ-specific, and stress-specific patterns of gene expression for nearly all of the Arabidopsis genes. These resources can thus provide clues to possible HRGP functions and/or allow researchers to focus their research projects. For example, in looking for phenotypic alterations in a HRGP mutant plant, microarray or MSPP data can guide the researcher in terms of the particular developmental times, organs, or conditions to examine in order to reveal a phenotype. Microarray and MPSS data are available for all but a few HRGPs. The majority of the AGP and EXT genes demonstrate organ-specific expression, while the remaining genes are expressed in multiple organs. Many AGPs, including classical AGPs, AG peptides, and at least one FLA, show pollen-specific expression. Likewise, root-specific AGPs are found in each AGP class. In contrast, pollen-specific expression of the EXT genes is restricted to the chimeric EXTs, most notably to certain LRXs (i.e. PEXs) and PERKs. Root-specific expression is exhibited by certain members of virtually all EXT classes. Approximately half of the PRPs show organ-specific expression, mostly in roots, while the rest are more widely expressed. Clearly, the notion that HRGPs in a particular class have some common organ-specific function appears unlikely, although the idea that certain AGPs are markers of cellular identity is supported by the organ-specific expression patterns revealed here (Knox et al., 1989). Comparing published northern and reverse transcription-PCR data on selected HRGP genes in studies conducted by various researchers with the microarray and MPSS data has consistently resulted in good agreement between these various methods to determine patterns of gene expression.

The recently updated Genevestigator Web site has considerably simplified the process of examining stress-induced gene expression in Arabidopsis microarrays. Virtually all HRGP genes are up- and down-regulated by various abiotic and biotic stress conditions. With the exception of some of the PRP genes, which exhibit common regulatory responses to auxin, zeatin, and infection by nematodes and *P. syringae*, it is difficult to summarize the diverse array of responses exhibited by the various HRGP genes. However, the coexpression database analysis takes into account these data, making common patterns of regulation much easier to recognize and examine. Nonetheless, if one is interested in a particular HRGP gene or in regulation by a particular stress condition, the data collected here constitute an ideal starting point for verification of this stress-induced gene regulation and for formulating functional hypotheses for particular HRGP genes.

HRGP Networks and Genes Involved in Posttranslational Modification

One unique genetic resource available to Arabidopsis researchers is the coexpression database. This database reports genes that are coexpressed with a gene of interest based on hundreds of different microarray gene analyses experiments. For HRGPs, this coexpression database offers the opportunity to reveal networks of genes associated with a given HRGP gene. In this study, the focus was placed on elucidating HRGP gene networks and in identifying candidate genes involved with the posttranslational modification of HRGPs, including genes involved with prolyl hydroxylation, glycosylation, and cross-linking. With regard to HRGP networks, it was remarkable that certain FLAs, namely FLA2, -7, -8, and -9, were coexpressed with so many different AGPs, EXTs, and PRPs. One interpretation of this result is that these FLAs play important roles in coordinating activities among various HRGP molecules; however, this and other interpretations must await functional characterization of these FLAs. Clearly, HRGP gene networks likely exist, given that sets of HRGP genes appear to be coregulated by a variety of conditions. It is possible that such regulatory networks are controlled by common regulatory sequences found in the HRGP genes. Efforts are currently under way as an extension of this work to identify such sequences using bioinformatics to allow for subsequent experimental testing of these elements and the transcription factors that bind to them.

It was hypothesized that a number of GT genes are expressed in conjunction with various HRGP genes to allow for the coordinated glycosylation of the encoded core protein. Furthermore, it was hypothesized that particular GTs would be responsible for synthesis of the various sugar linkages associated with the arabinogalactan polysaccharides attached to noncontiguous Hyp residues in AGPs, while other GTs would be associated with synthesis of the short arabinoside oligosaccharide chains attached to contiguous Hyp residues in EXTs and PRPs according to the Hyp continuity hypothesis (Tan et al., 2003). It was also hypothesized that GTs responsible for the addition of single Gal units to Ser residues in EXTs would be found. Moreover, based on the elucidated structures of dicot EXTs (Akiyama et al., 1980) and a well-characterized Hyp-AG isolated from transgenic tobacco (*Nicotiana tabacum*; Tan et al., 2004), and knowing the specificity of GTs, a minimum of 20 transferase activities are likely to be involved in the *O*-linked glycosylation of HRGPs. Specifically, for EXTs and PRPs, we predict one Ser-α-galactosyltransferase, at least one Hyp-β-arabinosyltransferase, one α-(1,2) arabinosyltransferase, and two β-(1,2)arabinosyltransferases, while for AGPs, we predict one Hyp-β-galactosyltransferase, one α-(1,5)arabinosyltransferase, at least four α-(1,3)arabinosyltransferases, at least three β-(1,3)galactosyltransferases, three β-(1,6)galactosyltransferases that add the three branch sites on the AG core, at least two β-(1,6)glucuronyltransferases, one α-(1,4)rhamnosyltransferase, and at least two α-(1,2) fucosyltransferases. Indeed, many GT genes are coexpressed with AGPs, EXTs, and PRPs. In fact, 36 different GTs representing 19 families were coexpressed with all three HRGP subfamilies, while some GTs are expressed only with two subfamilies or are restricted to one particular HRGP subfamily. While it is possible to speculate on the activities of these various GTs with respect to HRGPs based on their annotations and proposed mechanisms (i.e. inverting or retaining) in the CAZY database, such speculations would have to be tested by developing appropriate biochemical assays and/or obtaining and biochemically characterizing GT mutants. Indeed, such research is currently under way in a number of cell wall laboratories and is beginning to yield results. For example, it was recently shown that a mutant in the At2g35610 gene, encoding a GT77 family member, results in the production of underarabinosylated EXTs (Gille et al., 2009). Thus, the At2g35610 gene likely encodes one of the arabinosyltransferases required for EXT glycosylation and possibly for clustered

Hyp residues in certain AGPs, consistent with the identification of this gene in the coexpression data presented here in Tables VI and III, respectively.

Although only four plant P4Hs are cloned and characterized to date (two [P4H1 and P4H2] from Arabidopsis [Hieta and Myllyharju, 2002; Tiainen et al., 2005], one from tobacco [Yuasa et al., 2005], and one from *Chlamydomonas* [Keskiaho et al., 2007]), 13 P4H genes are predicted to exist for Arabidopsis (Vlad et al., 2007). The coexpression analysis performed here shows that only one of these P4H genes, namely P4H2, was consistently coexpressed with numerous HRGPs. This indicates that this P4H likely acts on AGPs, EXTs, and PRPs and is not restricted to a particular HRGP subfamily. Unfortunately, no published reports on P4H-2 mutants, or any P4H mutants in Arabidopsis, exist at present. However, the genetic redundancy in the P4H family may make such mutant work difficult. Nonetheless, a report that a P4H gene silenced by RNA interference in *Chlamydomonas* has an altered wall phenotype should bolster similar work in Arabidopsis (Keskiaho et al., 2007).

An acidic EXT peroxidase was isolated from tomato (*Solanum lycopersicum*) with EXT cross-linking activity (Schnabelrauch et al., 1996). It is also likely that PRPs and possibly AGPs undergo similar peroxidase-catalyzed cross-linking. In an effort to identify potential peroxidases involved with HRGP cross-linking, the coexpression database was used. Indeed, an acidic peroxidase (At3g03670) was identified using this approach and was coexpressed with the two most Tyr-rich EXTs. It will now be interesting to overexpress this enzyme for use in the EXT cross-linking assay and/or to obtain mutants in this gene and observe whether EXT is altered in these mutant plants in terms of more soluble EXTs, less cross-linked EXTs, or reduced amounts of the diisodityrosine/puchrescein cross-linking agent. It should be noted that several other peroxidase genes are also coexpressed and are worthy candidates for similar types of analysis.

HRGP Mutants Are Genetic Tools to Uncover HRGP Function

Genetic mutants are one of the most valuable resources available to the Arabidopsis community, as they provide insight to protein function and facilitate further research to elucidate the mechanism of action. This is clearly the case with HRGP research, where several genetic mutants in AGPs, EXTs, and PRPs are serving as useful tools to elucidate function. It should also be noted that for each informative HRGP mutant, there are many HRGP mutants that fail to reveal a phenotype. There are many potential reasons for such failure, including but not limited to one or more of the following: the existence of genetic redundancy or other genetic backup systems, the inability of certain mutants to adequately reduce mRNA or protein levels to reveal a phenotype, and the inability to examine the mutant under the proper environmental conditions to reveal its phenotype.

At present, several reports on HRGP mutants exist in Arabidopsis, including *agp17* (Gaspar et al., 2004), *agp18* (Acosta-Garcia and Vielle-Calzada, 2004), *agp19* (Yang et al., 2007), *sos5* (*fla4*; Shi et al., 2003), *agp30* (van Hengel and Roberts, 2003; van Hengel et al., 2004), *rsh-ext3* (Hall and Cannon, 2002), *lrx1* (Baumberger et al., 2001), and *perk13* (Humphrey et al., 2007). All these mutants have provided functional insights to the role of various AGPs and EXTs. The *agp17* mutant displays resistance to *Agrobacterium tumefaciens* transformation with reduced levels of AtAGP17 in the roots. An RNA interference approach was used to silence the AGP18 and reveal its role in female gametogenesis. An *agp19* mutant revealed that AGP19 plays a role in plant growth and development, specifically in cell division and expansion. Studies with the transposon-insertion mutant *agp30* suggest that AGP30I has a role in root regeneration and seed germination. The *sos5* mutant study indicates that FLA4 plays a role in cell expansion. The *rsh-ext3* mutant shows that EXT3 plays an important role in embryo development and cell plate formation, while the *lrx1* and *perk13* mutants indicate roles for LRX1 and PERK13 in root hair formation and root cell elongation, respectively.

There are currently 1,442 mutant lines available for nearly every HRGP gene, as shown in Tables II, V, and VIII and in Supplemental Tables S4, S7, and S10. While this list is now current, new mutant lines are continually being added to the collection, some of which are now being made available as homozygous knockout lines, saving the researcher valuable time and effort. In any event, once the mutant seed lines are received, they must be planted and verified by PCR analysis to confirm the presence of the mutation in the gene of interest. Mutations existing in the exon regions generally offer the highest probability of obtaining a null mutant and when available should probably be examined first. If a phenotype is observed in the mutant, it is important to confirm that the mutant phenotype is caused by the mutated gene of interest and not by another mutation elsewhere in the genome. Such confirmation can be achieved by studying other mutant lines (i.e. allelic mutants) for a gene of interest and observing the same mutant phenotype or by complementing the original mutant with the wild-type version of the gene of interest. Although mutants affecting the HRGP core proteins allow for the assessment of a particular HRGP's functional role, obtaining mutants in the genes responsible for HRGP posttranslational modification (i.e. GTs, P4Hs, peroxidases) offers perhaps even greater opportunities to address and reveal HRGP function, as multiple HRGPs would be affected by such a mutation.

CONCLUSION

The BIO OHIO bioinformatics program reported here represents a valuable tool to mine genomic databases for HRGP genes, including AGPs, EXTs, PRPs, chimeric HRGPs, and hybrid HRGPs. While this program was utilized to mine the Arabidopsis proteome, it can now be utilized to examine proteomes resulting from other plant genome projects, namely poplar (*Populus* species), rice (*Oryza sativa*), *Physcomitrella*, and *Chlamydomonas*. Preliminary evidence indicates, not surprisingly, that poplar is most similar to Arabidopsis in terms of its HRGP inventory, while the other species have considerable differences from the dicot HRGP inventory. In Arabidopsis, there are many surprises with respect to the HRGP family members beyond just finding new putative HRGPs, including finding HRGPs that apparently lack signal peptides, the predicted existence of GPI anchor addition sequences in certain EXTs, the numerous HRGPs that show organ-specific expression, and the likely existence of coregulated HRGP networks. Depending upon an investigator's interest, there is now a wealth of information provided to guide future HRGP research. Many of these predictions will require verification or confirmation, but hypotheses can now be formed and specific experiments designed based on the information presented here to facilitate future HRGP research.

Refinements to the BIO OHIO program are possible. In particular, reducing the number of false positives during a search and improving or developing search strategies to identify the chimeric HRGPs, particularly chimeric AGPs and chimeric PRPs, represent two of the most challenging areas for improving the predictive power of the program. In addition to the sliding windows approach, other more novel approaches are being examined to improve the predictive power of the program, including using hidden Markov models, neural networks, as well as supervised and unsupervised learning approaches.

Finally, while the program was specifically developed to identify HRGPs from plant genomic data, it can be readily adapted to identify other proteins or protein families. The ability to select any amino acid bias or sequence motif of interest should make this program attractive to other researchers, including those outside of the plant community, who wish to screen whole genome protein sequences meeting their desired criteria. In addition, this program can be used to screen virtually any protein database, including those created manually or from EST databases.

MATERIALS AND METHODS

Development and Basic Operation of the BIO OHIO Bioinformatics Program

A Perl program, named BIO OHIO, was written that analyzes each predicted protein sequence in the Arabidopsis (*Arabidopsis thaliana*) genome. This program is available upon request along with a user manual describing the use and operation of this program; however, an abbreviated version of the program is accessible at http://132.235.14.51/functional_genomics.html. The database used (i.e. ATH1.pep) was dated June 10, 2004, and downloaded from The Institute for Genomic Research (ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/). The program is able to categorize proteins based on various characteristics and patterns of amino acids as specified by the user/researcher. For each identified protein or "hit," the following information was provided: (1) the Arabidopsis Genome Initiative locus identifier and sequence name; (2) the entire protein sequence; (3) the length of the protein; (4) the total PAST percentage for each protein; (5) analysis for the presence of a signal peptide within the first 50 amino acid residues; and (6) analysis for the presence of a GPI anchor addition sequence. In addition, the program provided analysis of repeated sequences within the proteins. In particular, the presence of AP, PA, SP, and TP dipeptide repeats were noted, as these sequences are typically associated with known AGPs. Protein hits were classified as AGPs if they did not contain repeats associated with EXTs or PRPs (e.g. multiple SP_4, SP_3, or PPV repeats) but contained predominantly AP, PA, SP, or TP repeats. In order to verify the predictions easily, the program predicted signal peptides and GPI anchor addition sequences and also allowed direct connection to the SignalP Web site (http://www.cbs.dtu.dk/services/SignalP/) to verify signal peptides, the Plant big-PI predictor Web site (http://mendel.imp.ac.at/gpi/plant_server.html) to verify GPI anchor predictions, and the TAIR Web site (http://arabidopsis.org/) for gene and protein information. When conflicts arose between BIO OHIO and the SignalP Web site or the Plant big-PI predictor Web site, data from the SignalP Web site or the Plant big-PI predictor Web site were used.

Finding Classical AGPs and AG Peptides Using Biased Amino Acid Compositions and Finding FLAs by Searching for Fasciclin Motifs

Classical AGPs were identified as proteins of any length that consisted of 50% or greater of the amino acids P, A, S, and T (PAST). AG peptides were identified as proteins of 50 to 90 amino acids in length consisting of 35% or greater PAST. A reduced PAST level was used, since AG peptides usually contain an N-terminal signal peptide and possibly a C-terminal GPI anchor addition signal sequence, which can make up about half of the peptide and contain little PAST. FLAs were designated as proteins containing the consensus motif [MALIT]T[VILS][FLCM][CAVT][PVLIS][GSTKRNDPEIV]+[DNS][DSENAGE]+[ASQM]. This motif was constructed by comparison of all known Arabidopsis FLAs as reported by Johnson et al. (2003a).

Finding EXTs by Searching for SP_4 and SP_3 Repeat Motifs

The program allowed for searches of any given amino acid string written as a regular expression. Thus, EXTs were identified by searching for the occurrence of two or more SP_4 (or SP_3) repeats in the protein. Since some of these hits were already annotated as PERKs in the TAIR database, we also manually included other known members of this family from the published literature (Baumberger et al., 2003a; Nakhamchik et al., 2004). Hits were examined for the location and distribution of SP_4 and SP_3 repeats as well as for the occurrence of other repeating sequences, including YXY. In addition, these sequences were examined for potential signal peptides and GPI anchor addition sequences as described above.

Finding PRPs by Using Biased Amino Acid Compositions and by Searching for PPVX(K/T) and KKPCPP Repeat Motifs

PRPs were first identified by searching for a biased amino acid composition of greater then 45% PVKCYT (Fowler et al., 1999). PRPs were also identified by searching for the occurrence of two or more PPVX(K/T) (where X represents any amino acid) and KKPCPP motifs (Fowler et al., 1999). Hits were examined for the location and distribution of these repeats as well as PPV repeat units. In addition, these sequences were examined for potential signal peptides and GPI anchor addition sequences as described above.

Finding Amino Acid Sequence Repeats in a Protein Sequence

Operating on a Bio::Perl sequence object, a frequency function determines the repeating elements in a given protein sequence. The length of the repeating elements is a parameter that can be set by specifying a minimum length of an element and a maximum length of an element. This variability allows a very thorough examination of the sequence. For each length that lies between the minimum and maximum length, set in the parameters, a sliding window of that length is used and shifted across the sequence, in increments of one amino acid, starting at position 1 and ending at the last position: the length of the sliding window + 1. The discovered elements are stored in a hash structure, with the subsequence of the sliding window as the key and the number of occurrences as the entry. Upon this hash structure, the percentages are computed and stored. This extended hash is then passed onto a visualization function that adds html tags around a currently highlighted pattern and thus allows the analysis of pattern distribution among the complete amino acid sequence.

Searching User-Defined Regions (Sliding Windows) to Find HRGP Domains in a Protein Sequence

The sliding window is a feature built into the BIO OHIO program that can be used for looking at small sections of a protein rather than the protein as a whole. The sliding window starts at the beginning of the protein and slides along the sequence, searching for a biased amino acid composition in a user-designated window size. The sliding windows feature is most useful to find chimeric HRGPs, since only small sections of these proteins contain HRGP motifs. The sliding window can also be used to visualize HRGP regions in proteins found using other searches, as with FLAs or PAGs.

Annotation of Examined Sequences following Our Analysis

Another feature of the program is the ability to create custom annotations for genes identified following a search. This option takes the form of a box into which one types particular keywords about the identified gene. Once the keywords are entered for a particular gene, that gene will appear with an asterisk in all future searches as an indicator that it was identified previously. The keywords are also searchable so that the custom-annotated genes can easily be found at a later time.

Finding Potential HRGPs by Searching Annotations in the Arabidopsis Database

In addition to using biased amino acid composition and repeat searches, an annotation search feature built into the BIO OHIO program was also utilized. Keywords, including extensin, Pro-rich, arabinogalactan, plastocyanin, and Hyp, were entered to see if any additional proteins in the database were already annotated with these keywords. These proteins were then examined as described above to determine whether they were indeed likely AGPs, EXTs, or PRPs.

BLAST Analysis

BLAST analysis was performed on each identified HRGP using TAIR WU-Blast 2.0 (http://www.arabidopsis.org/wublast/index2.jsp) to identify other potential HRGP sequences and to provide insight to HRGP sequences with the greatest similarity. Specifically, the BLASTX: NT query to AA db was used along with the AGI Proteins (Protein) database. BLAST searches were conducted with the "filter query" option both on and off.

Elucidation of Expression Patterns of HRGP Genes Using Public Databases

In order to elucidate patterns of gene expression for the predicted HRGPs, three public databases were searched: Genevestigator (https://www.genevestigator.ethz.ch/), Arabidopsis Membrane Protein Library (http://www.cbs.umn.edu/arabidopsis/), and Arabidopsis MPSS Plus Database (http://mpss.udel.edu/at/).

Identification of HRGP, GT, P4H, and Peroxidase Genes Coexpressed with the Predicted HRGP Genes in Arabidopsis

All HRGP genes were examined with respect to coexpressed genes using The Arabidopsis Co-Response Database (http://csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html). At this site, "single gene query" was selected. Each of the HRGPs was searched using the four different matrices: nasc0271, atge0100, atge0200, and atge0250. The default settings for coefficient and output were used. These results were examined, and only GTs, P4Hs, peroxidases, and other HRGPs that were coexpressed with a given HRGP were selected.

Identification of Gene Structure and Genetic Mutants for the Identified HRGP Genes

Information on HRGP gene structures was obtained from the TAIR database (http://www.arabidopsis.org). In order to determine if genetic mutants exist in each of these predicted HRGP genes, T-DNAexpress: The SIGnAL Arabidopsis Gene Mapping Tool (http://signal.salk.edu/cgi-bin/tdnaexpress) was utilized. All reported mutant lines were documented following the search.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Amino acid sequences of AGPs.

Supplemental Figure S2. Genevestigator anatomy expression for AGPs.

Supplemental Figure S3. Genevestigator stimulus expression for AGPs.

Supplemental Figure S4. Arabidopsis Membrane Protein Library data for AGPs.

Supplemental Figure S5. MPSS data for AGPs.

Supplemental Figure S6. Amino acid sequences of EXTs.

Supplemental Figure S7. Genevestigator anatomy expression for EXTs and hybrid HRGPs.

Supplemental Figure S8. Genevestigator stimulus expression for EXTs and hybrid HRGPs.

Supplemental Figure S9. Arabidopsis Membrane Protein Library data for EXTs and hybrid HRGPs.

Supplemental Figure S10. MPSS data for EXTs and hybrid HRGPs.

Supplemental Figure S11. Amino acid sequences of PRPs.

Supplemental Figure S12. Genevestigator anatomy expression for PRPs.

Supplemental Figure S13. Genevestigator stimulus expression for PRPs.

Supplemental Figure S14. Arabidopsis Membrane Protein Library data for PRPs.

Supplemental Figure S15. MPSS data for PRPs.

Supplemental Table S1. Results of HRGP BLAST searches with filter on (worksheet A) and off (worksheet B).

Supplemental Table S2. HRGPs, GTs, P4Hs, and peroxidases coexpressed with AGPs.

Supplemental Table S3. Locations of introns and exons in AGPs.

Supplemental Table S4. Locations of available T-DNA mutant lines for AGPs.

Supplemental Table S5. HRGPs, GTs, P4Hs, and peroxidases coexpressed with EXTs.

Supplemental Table S6. Locations of introns and exons in EXTs and hybrid HRGPs.

Supplemental Table S7. Locations of available T-DNA mutant lines for EXTs and hybrid HRGPs.

Supplemental Table S8. HRGPs, GTs, P4Hs, and peroxidases coexpressed with PRPs.

Supplemental Table S9. Locations of introns and exons in PRPs.

Supplemental Table S10. Locations of available T-DNA mutant lines for PRPs.

Received March 25, 2010; accepted April 12, 2010; published April 15, 2010.

LITERATURE CITED

Acosta-Garcia G, Vielle-Calzada JP (2004) A classical arabinogalactan protein is essential for the initiation of female gametogenesis in *Arabidopsis*. Plant Cell **16:** 2614–2628

Akiyama Y, Mori M, Kato K (1980) ^{13}C-NMR analysis of hydroxyproline arabinosides from *Nicotiana tabacum*. Agric Biol Chem **44:** 2487–2489

Baldwin TC, Domingo C, Schindler T, Seetharaman G, Stacey N, Roberts K (2001) DcAGP1, a secreted arabinogalactan protein, is related to a family of basic proline-rich proteins. Plant Mol Biol **45:** 421–435

Baumberger N, Doesseger B, Guyot R, Diet A, Parsons R, Clark M, Simmons MP, Bedinger P, Goff S, Ringli C, et al (2003a) Whole-genome comparison of leucine-rich repeat extensins in Arabidopsis and rice: a conserved family of cell wall proteins form a vegetative and a reproductive clade. Plant Physiol **131:** 1313–1326

Baumberger N, Ringli C, Keller B (2001) The chimeric leucine-rich repeat/extensin cell wall protein LRX1 is required for root hair morphogenesis in *Arabidopsis thaliana*. Genes Dev **15:** 1128–1139

Baumberger N, Steiner M, Ryser U, Keller B, Ringli C (2003b) Synergistic interaction of the two paralogous Arabidopsis genes LRX1 and LRX2 in cell wall formation during root hair development. Plant J **35:** 71–81

Borner GHH, Lilley KS, Stevens TJ, Dupree P (2003) Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis: a proteomic and genomic analysis. Plant Physiol **132:** 568–577

Brady JD, Sadler IH, Fry SC (1996) Di-isodityrosine, a novel tetrameric derivative of tyrosine in plant cell wall proteins: a new potential cross-link. Biochem J **315:** 323–327

Brady JD, Sadler IH, Fry SC (1998) Pulcherosine, an oxidatively coupled trimer of tyrosine in plant cell walls: its role in cross-link formation. Phytochemistry **47:** 349–353

Cannon MC, Terneus K, Hall Q, Tan L, Wang Y, Wegenhart BL, Chen L, Lamport DTA, Chen Y, Kieliszewski MJ (2008) Self-assembly of the plant cell wall requires an extensin scaffold. Proc Natl Acad Sci USA **105:** 2226–2231

Cassab GI (1998) Plant cell wall proteins. Annu Rev Plant Physiol Plant Mol Biol **49:** 281–309

Fowler TJ, Bernhardt C, Tierney ML (1999) Characterization and expression of four proline-rich cell wall protein genes in Arabidopsis encoding two distinct subsets of multiple domain proteins. Plant Physiol **121:** 1081–1091

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

Gille S, Hänsel U, Ziemann M, Pauly M (2009) Identification of plant cell wall mutants by means of a forward chemical genetic approach using hydrolases. Proc Natl Acad Sci USA **106:** 14699–14704

Hall Q, Cannon M (2002) The cell wall hydroxyproline-rich glycoprotein RSH is essential for normal embryo development in *Arabidopsis*. Plant Cell **14:** 1161–1172

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

Hieta R, Myllyharju J (2002) Cloning and characterization of a low molecular weight prolyl 4-hydroxylase from *Arabidopsis thaliana*: effective hydroxylation of proline-rich, collagen-like, and hypoxia-inducible transcription factor alpha-like peptides. J Biol Chem **277:** 23965–23971

Humphrey TV, Bonetta DT, Goring DR (2007) Sentinels at the wall: cell wall receptors and sensors. New Phytol **176:** 7–21

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

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

José-Estanyol M, Puigdomènech P (2000) Plant cell wall glycoproteins and their genes. Plant Physiol Biochem **38:** 97–108

Keskiaho K, Hieta R, Sormunen R, Myllyharju J (2007) *Chlamydomonas reinhardtii* has multiple prolyl 4-hydroxylases, one of which is essential for proper cell wall assembly. Plant Cell **19:** 256–269

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

Kjellbom P, Snogerup L, Stohr C, Reuzeau C, McCabe PF, Pennell RI (1997) Oxidative cross-linking of plasma membrane arabinogalactan proteins. Plant J **12:** 1189–1196

Knox J, Day S, Roberts K (1989) A set of cell surface glycoproteins forms an early position, but not cell type, in the root apical carota L. Development **106:** 47–56

Lee S, Saravanan RS, Damasceno CMB, Yamane H, Kim B, Rose JKC (2004) Digging deeper into the plant cell wall proteome. Plant Physiol Biochem **42:** 979–988

Liu C, Mehdy M (2007) A nonclassical arabinogalactan protein gene highly expressed in vascular tissues, AGP31, is transcriptionally repressed by methyl jasmonic acid in Arabidopsis. Plant Physiol **145:** 863–874

Merkouropoulos G, Barnett DC, Shirsat AH (1999) The Arabidopsis extensin gene is developmentally regulated, is induced by wounding, methyl jasmonate, abscisic, and salicylic acid and codes for a protein with unusual motifs. Planta **208:** 212–219

Nakhamchik A, Zhao Z, Provart NJ, Shiu S, Keatley SK, Cameron RK, Goring DR (2004) A comprehensive expression analysis of the Arabidopsis proline-rich extensin-like receptor kinase gene family using bioinformatic and experimental approaches. Plant Cell Physiol **45:** 1875–1881

Nickel W (2003) The mystery of nonclassical protein secretion: a current view on cargo proteins and potential export routes. Eur J Biochem **270:** 2109–2119

Nothnagel EA (1997) Proteoglycans and related components in plant cells. Int Rev Cytol **174:** 195–291

Schnabelrauch LS, Kieliszewski MJ, Upham BL, Alizedeh H, Lamport DTA (1996) Isolation of pI 4.6 extensin peroxidase from tomato cell suspension cultures and identification of Val-Tyr-Lys as putative intermolecular cross-link site. Plant J **9:** 477–489

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

Schultz CJ, Johnson KL, Currie G, Bacic A (2000) The classical arabinogalactan protein gene family of *Arabidopsis*. Plant Cell **12:** 1751–1768

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

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

Sherrier DJ, Prime TA, Dupree P (1999) Glycosylphosphatidylinositol-anchored cell-surface proteins from Arabidopsis. Electrophoresis **20:** 2027–2035

Shi H, Kim Y, Guo Y, Stevenson B, Zhu JK (2003) The *Arabidopsis* SOS5 locus encodes a putative cell surface adhesion protein and is required for normal cell expansion. Plant Cell **15:** 19–32

Showalter AM (1993) Structure and function of plant cell wall proteins. Plant Cell **5:** 9–23

Showalter AM (2001) Arabinogalactan-proteins: structure, expression and function. Cell Mol Life Sci **58:** 1399–1417

Sun W, Xu J, Yang J, Kieliszewski MJ, Showalter AM (2005) The lysine-rich arabinogalactan-protein subfamily in Arabidopsis: gene expression, glycoprotein purification and biochemical characterization. Plant Cell Physiol **46:** 975–984

Svetek J, Yadav MP, Nothnagel EA (1999) Presence of a glycosylphosphatidylinositol lipid anchor on rose arabinogalactan proteins. J Biol Chem **274:** 14724–14733

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

Tan L, Qiu F, Lamport DTA, Kieliszewski MJ (2004) Structure of a hydroxyproline (Hyp)-arabinogalactan polysaccharide from repetitive Ala-Hyp expressed in transgenic *Nicotiana tabacum*. J Biol Chem **279:** 13156–13165

Tiainen P, Myllyharju J, Koivunen P (2005) Characterization of a second *Arabidopsis thaliana* prolyl 4-hydroxylase with distinct substrate specificity. J Biol Chem **280:** 1142–1148

van Hengel AJ, Barber C, Roberts K (2004) The expression patterns of arabinogalactan-protein AtAGP30 and GLABRA2 reveal a role for abscisic acid in the early stages of root epidermal patterning. Plant J **39:** 70–83

van Hengel AJ, Roberts K (2003) AtAGP30, an arabinogalactan-protein in the cell walls of the primary root, plays a role in root regeneration and seed germination. Plant J **36:** 256–270

Vlad F, Spano T, Vlad D, Daher FB, Ouelhadj A, Kalaitzis P (2007) Arabidopsis prolyl 4-hydroxylases are differentially expressed in response to hypoxia, anoxia and mechanical wounding. Physiol Plant **130:** 471–483

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

Yoshiba Y, Aoki C, Iuchi S, Nanjo T, Seki M, Sekiguchi F, Yamaguchi-Shinozaki K, Shinozaki K (2001) Characterization of four extensin genes in *Arabidopsis thaliana* by differential gene expression under stress and non-stress conditions. DNA Res **8:** 115–122

Youl JJ, Bacic A, Oxley D (1998) Arabinogalactan-proteins from *Nicotiana alata* and *Pyrus communis* contain glycosylphosphatidylinositol membrane anchors. Proc Natl Acad Sci USA **95:** 7921–7926

Yuasa K, Toyooka K, Fukuda H, Matsuoka K (2005) Membrane-anchored prolyl hydroxylase with an export signal from the endoplasmic reticulum. Plant J **41:** 81–94