Genome-wide identification, classification, and expression analysis of the phytocyanin gene family in *Phalaenopsis equestris*

L. XU1,2, X.J. WANG1, T. WANG1, and L.B. LI1*

State Key Laboratory of Forest Genetics and Tree Breeding, Key Laboratory of Silviculture of the State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, P.R. China

1 College of Horticulture and Landscape, Hunan Agricultural University, Changsha 410128, P.R. China

Abstract

Phytocyanins (PCs) are ancient blue copper-binding proteins in plants that bind to single type I copper atoms and function as electron transporters. PCs play an important role in plant development and stress resistance. Many PCs are considered to be chimeric arabinogalactan proteins (AGPs). Previously, 38, 62, and 84 PC genes were identified in *Arabidopsis thaliana*, *Oryza sativa*, and *Brassica rapa*, respectively. In this study, we identified 30 putative PC genes in the orchid *Phalaenopsis equestris* through comprehensive bioinformatics analysis. Based on phylogeny and motif constitution, the *P. equestris* phytocyanins (PePCs) were divided into five subclasses: 10 early nodulin-like proteins, 10 uclacyanin-like proteins, five stellacyanin-like proteins, four plantacyanin-like proteins, and one unknown protein. Structural and glycosylation predictions suggested that 16 PePCs were glycosylphosphatidylinositol-anchored proteins localized to the plasma membrane, 22 PePCs contain N-glycosylation sites, and 14 are chimeric AGPs. Phylogenetic analysis indicated that each subfamily was derived from a common ancestor before the divergence of monocot and dicot lineages and that the expansion of the PC subfamilies occurred after the divergence of orchids and *Arabidopsis*. The number of exons in PC genes was conserved. Expression analysis in four tissues revealed that nine PC genes were highly expressed in flowers, stems, and roots, suggesting that these genes play important roles in growth and development in *P. equestris*. The results of this study lay the foundation for further analysis of the functions of this gene family in plants.

Additional key words: arabinogalactan proteins, early nodulin-like proteins, plantacyanin-like proteins, stellacyanin-like proteins, uclacyanin-like proteins.

Introduction

Phytocyanins (PCs) are ancient plant blue copper-binding proteins (BCPs) that function as electron transporters and bind to single type I copper atoms (Rydén and Hunt 1993, De Rienzo et al. 2000, Ruan et al. 2011). All PCs have an eight-stranded Greek key β-barrel or β-sandwich fold and two conserved disulfide-bridged cystein (Cys) residues, and some possess four copper ligands: two histidine (His), one Cys, and one methionine (Met) or glutamine (Gln) (Garrett et al. 1984, Hart et al. 1996). Comprehensive bioinformatics analysis revealed that 38, 62, and 84 PC genes are present in *Arabidopsis thaliana*, *Oryza sativa*, and *Brassica rapa*, respectively (Mashiguchi et al. 2009, Ma et al. 2011, Li et al. 2013). The PCs can be divided into four subfamilies based on the characteristics of their copper ligand residues, spectroscopic and redox properties, and the domain organization of the protein: uclacyanins (UCs), stellacyanins (SCs), plantacyanins (PLCs), and early nodulin-like proteins (ENODLs) (Nersissian et al. 1998, Mashiguchi et al. 2004). All PCs possess a plastocyanin-like domain (PCLD). UCs and PCs contain the same copper ligand residues (two His, one Cys and one Met) in their domains. However, the

Submitted 11 December 2015, last revision 31 March 2016, accepted 19 April 2016.

Abbreviations: AG - arabinogalactan; AGPs - arabinogalactan proteins; ALR - arabinogalactan-like region; BCPs - blue copper-binding proteins; BLASTP - protein basic local alignment search tool; ENODLs - early nodulin-like proteins; GAS - glycosylphosphatidylinositol-anchor signal; GPI - glycosylphosphatidylinositol; HRGPs - hydroxyproline-rich glycoproteins; PCs - phytocyanins; PCLD - plastocyanin-like domain; PLAs - phytocyanin-like arabinogalactan proteins; PLCs - plantacyanins; PLCLs - plantacyanin-like proteins; RPKM - reads per kilobase per million mapped reads; SCs - stellacyanins; SCLs - stellacyanin-like proteins; SP - signal peptide; UCs - uclacyanins; UCLs - uclacyanin-like proteins.

Acknowledgments: This work was supported by grants from the Ministry of Science and Technology (No. 2012BAD01B0702 and No. 2013AA102607).

* Corresponding author: fax: (+86)01062888687, e-mail: lilubin@126.com
UCs are chimeric glycoproteins, whereas PLCs are not (Nersissian et al. 1998). Similarly, SCs have both a copper-binding domain (two His, one Cys, and one Gln) and a glycoprotein-like domain (Mann et al. 1992, Van Driessche et al. 1995). SCs and UCs contain N-linked glycosylation sites through an asparagine (Asn) residue, as well as O-linked glycosylation sites through serine (Ser) and hydroxyproline (Hyp) residues. The structures of ENODLs are similar to those of the other three subfamilies, whereas ENODLs lack amino acid residues for copper binding, and the vast majority of ENODLs in Arabidopsis and rice are chimeric arabinogalactan proteins (AGPs) (Greene et al. 1998, Mashiguchi et al. 2004, 2009, Ma et al. 2011).

Arabinogalactan proteins (AGPs) are a highly diverse group of cell surface glycoproteins containing type II AG polysaccharide chains attached to the core protein backbone by O-linked glycosylation, which usually comprise > 90% of the total molecular mass (Seifert and Roberts 2007). AGPs are classified into classical or nonclassical types, i.e., AG peptides (10 to 15 amino acid residues) and fasciclin-like AGPs (Gaspar et al. 2001, Schultz et al. 2002). The arabinogalactosylated domains and different proline- or Hyp-rich glycoprotein motifs determine the type of AGP. In addition to possessing at least one arabinogalactosylated domain, chimeric AGPs also contain a domain with an unrelated motif. The PC gene family belongs to the AGP superfamily, and bioinformatics analysis revealed that 29 of 38 AtPCs and 38 of 62 OsPCs are putative chimeric AGPs (Ma et al. 2011). AGPs are involved in various biological processes, such as plant cell growth and development, pollen tube elongation, and plant-microbe interactions (Gaspar et al. 2001, Seifert and Roberts 2007, Tan et al. 2012).

PCs were recently shown to play an important role in plant development and stress resistance. PCs participate in the plant defense responses to abiotic stresses, such as drought, salt, cold, and metal ion stress (Ozturk et al. 2002, Diab et al. 2004, Ezaki et al. 2005, Ma et al. 2011, Wu et al. 2011). The UC gene subfamily is strongly expressed in roots and stems and is highly conserved in the dicotyledon B. rapa and the monocotyledon O. sativa, suggesting that this subfamily has played important roles in polyploid crops throughout evolution (Ma et al. 2011, Li et al. 2013). SC genes are mainly expressed in roots and inflorescences in rice and induced by aluminum stress and oxidative stress in Arabidopsis (Richards et al. 1998, Ezaki et al. 2000). Similarly, PLC genes are also stress-related and are highly expressed in inflorescence tissue, especially in the transmitting tract of the pistil (Kreps et al. 2002, Dong et al. 2005). Notably, PLCs function as signaling molecules, as do ENODLs, which have node-specific expression patterns in legumes and are thought to be involved in defense responses (Nersissian et al. 1998, Fedorova et al. 2002). ENODLs may also function in organ development, as several ENODL genes are expressed in non-legumes, including the apical buds of Pharbitis nil and floral organs of Arabidopsis (Yoshizaki et al. 2000, Mashiguchi et al. 2009). Recently, the ENODL gene SvNod1, which encodes an early nodulin-like protein, was found to be specifically expressed in mycorrhizal orchid (Serapias vomeracea) protocorms (Perotto et al. 2014), suggesting that ENODLs play important roles not only in the noduleation process, but also in symbiotic development.

The genome sequence of the orchid Phalaenopsis equestris is an important resource that can be used to identify and bioinformatically analyze putative PCs in this commercially important species. The aim of this study was to identify genes encoding putative PCs in the P. equestris genome, phylogenetic analysis of the PePCs, and determination of the amino acid sequences of these proteins. In addition, we tried to analyze the exon-intron structures and expression of PePC genes in various tissues. The results of this study can provide a basis for further elucidating the functions of this gene family in plants, especially the functions of PePC genes during plant-fungus interactions in orchid mycorrhhizae.

Materials and methods

Identification of PePCs and bioinformatics analysis: All protein files from P. equestris (pep files) were downloaded from the Phalaenopsis equestris genome database (ftp://ftp.genomics.org.cn/from_BGISZ/20130120/). The PePC family members were identified with the BLAST tool using AtPC amino acid sequences (Ma et al. 2011) as a query against the P. equestris database. Default parameters were used in the BLAST analyses, and false hits (E-value < 0.01) were removed by manual inspection. The simple modular architecture research tool (SMART, http://smart.embl-heidelberg.de) and a database of protein domains, families and functional sites (PROSITE, http://prosite.expasy.org) were used to confirm the existence of the PCLD.

The N-terminal signal peptide (SP) of each PePC was examined using SignalP 4.1 (Petersen et al. 2011). C-terminal GPI-anchor signals (GASs) were predicted using the big-PI plant predictor (http://mendel.imp.ac.at/gpi/plant_server.html) and protein subcellular localization prediction tool PSORT (http://psort.hgc.jp/form.html). N-glycosylation sites were predicted using NetNGlyc 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/). Putative AG glycomodules in the AG-like region (ALR) were predicted following previously described criteria (Shpak et al. 2001, Schultz et al. 2002, 2004, Tan et al. 2003, Estévez et al. 2006, Mashiguchi et al. 2009, Ma and Zhao 2010, Showalter et al. 2010). Clustered noncontiguous Pro residue motifs ([Ala/Ser/Thr/Gly]-Pro-[X(0,10)-[Ala/Ser/Thr/Gly]-Pro] and contiguous Pro residue motifs ([Ala/Ser/Thr/Gly]-Pro) were defined as putative AG glycomodules, and Ser-Pro2 motif were defined as putative extensin glycomodules.
Multiple sequence alignment of PCLDs and phylogenetic analysis: Multiple alignments of the amino acid sequences of the PCLDs in the PePCs were performed using Clustal X 2.1 (Larkin et al. 2007). The alignment results were manually edited according to the characteristics of the domain.

All PePC and AtPC amino acid sequences were aligned using Clustal X 2.1. The resulting sequence alignment was converted into MEGA format for tree construction by the neighbor-joining method using MEGA 6.06 (Tamura et al. 2013). Bootstrap analysis was performed with 1 000 replicates.

The gene structures were retrieved from the *P. equestris* genome database. The exon-intron structures of *P. equestris* PC genes were mapped using the gene structure display server (GSDS, http://gsds1.cbi.pku.edu.cn) (Guo et al. 2007).

Expression analysis of PePC genes: The expression data for all genes in four tissues (flowers, leaves, stems, and roots) of *P. equestris* were obtained from the genome sequence of the orchid *Phalaenopsis equestris* (http://www.nature.com/ng/journal/v47/n1/full/ng.3149.html) (Cao et al. 2015). A heatmap of PC gene expression clusters from each tissue was constructed using HemI 1.0 (Deng et al. 2014), and the expression values of PC genes were presented as RPKM (reads per kilobase per million mapped reads)-normalized log2 transformed counts.

Results

We initially identified 47 PC-like proteins in the *P. equestris* protein database using the BLASTP search tool. After verifying the PCLDs of these proteins, 17 of the proteins were found to lack PCLDs and were therefore eliminated from further analysis. One protein (PEQU_10123) lacked a PCLD according to SMART analysis, but it contained a PCLD based on PROSITE. Since this protein contains characteristic conserved amino acid sequences of PCs (two disulfide-bridged Cys residues and four copper ligands), we consider it to represent a PePC (Fig. 1 Suppl.). We ultimately identified a total of 30 genes encoding PePCs in the *P. equestris* genome. The *Arabidopsis* genome contains 38 PC genes, which is very similar to the number of PC genes in this orchid. By contrast, the rice and Chinese cabbage genomes contain 62 and 84 PC genes, respectively, indicating that rice has more than twice as many PC genes as orchid and Chinese cabbage has approximately three-times as many PC genes.

Multiple sequence alignment revealed that the Cys residues involved in the formation of disulfide linkages to maintain the stability of PCLD structure are highly conserved in these PePCs (Fig. 1 Suppl.), which is consistent with the structures of AtPCs, OsPCs and BrPCs. Nineteen PePCs contain four copper ligand residues to bind copper (Fig. 1 Suppl.). Based on the types of copper ligand residues and the absence of glycosylation sites on their proteins backbones, these PePCs were classified into three subfamilies: 10 uclacyanin-like proteins (UCLs), five stellacyanin-like proteins (SCLs), and four plantacyanin-like proteins (PLCLs; Fig. 1 Suppl., Table 1 Suppl.). The four PePLCLs are in a separate group from the PeUCLs due to the absence of glycosylation sites on their protein backbones (Table 1 Suppl.). The remaining 11 PePCs without copper ligand residues were classified into two subfamilies: 10 PeENODLs and one unknown PC-like protein (PEQU_09781; Fig. 1 Suppl., Table 1 Suppl.). We predicted N-terminal SP, C-terminal GAS, and N-glycosylation sites in the PePC protein backbones to more clearly elucidate their structural characteristics. Based on this analysis, 21 PePCs were predicted to possess N-terminal SPs responsible for conventional protein secretion. We combined two algorithms, finding that 16 PePCs may have GAS sites (Table 1 Suppl.). All GASs

![Fig. 1. Schematic representation of six groups of PePCs. The diagram of features of PePC domains was generated with MyDomains (http://prosite.expasy.org/cgi-bin/prosite/mydomains/). SP - signal peptide, PCLD - plastocyanin-like domain, ALR - AGP-like region, GAS - GPI-anchor signal, AGPs - arabinogalactan proteins. The figure is not drawn to scale.](http://example.com/fig1.png)
Fig. 2. Phylogenetic tree based on amino acid sequences of AtPCs and PePCs, including AtENODL1 (At5g53870), AtENODL2 (At4g27520), AtENODL3 (At4g28365), AtENODL4 (At4g32490), AtENODL5 (At3g18590), AtENODL6 (At1g48940), AtENODL7 (At1g79800), AtENODL8 (At1g64640), AtENODL9 (At3g20570), AtENODL10 (At5g57920), AtENODL11 (At2g23990), AtENODL12 (At4g30590), AtENODL13 (At5g25900), AtENODL14 (At2g25060), AtENODL15 (At4g31840), AtENODL16 (At3g01070), AtENODL17 (At5g15350), AtENODL18 (At1g08500), AtENODL19 (At4g12880), AtENODL20 (At2g27035), AtENODL21 (At5g14345), AtENODL22 (At1g17800), AtUC1 (At1g22480), AtUC2 (At1g22320), AtUC3 (At2g32300), AtUC4 (At2g44790), AtUC5 (At3g27200), AtUC6 (At3g60270), AtUC7 (At3g60280), AtUC8 (At5g07475), AtSCI (At2g26720), AtSC2 (At2g31050), AtSC3 (At5g20230), AtSC4 (At5g26330), AtPLCL1 (At2g02850), At1g45063, At3g53330, and At4g01380 from Arabidopsis thaliana; PeENODL1 (PEQU_00681), PeENODL2 (PEQU_01525), PeENODL3 (PEQU_07653), PeENODL4 (PEQU_08393), PeENODL5 (PEQU_08610), PeENODL6 (PEQU_11121), PeENODL7 (PEQU_12079), PeENODL8 (PEQU_13421), PeENODL9 (PEQU_15783), PeENODL10 (PEQU_36125), PeUC1 (PEQU_00112), PeUC2 (PEQU_00113), PeUC3 (PEQU_06004), PeUC4 (PEQU_06005), PeUC5 (PEQU_07626), PeUC6 (PEQU_10123), PeUC7 (PEQU_14669), PeUC8 (PEQU_15884), PeUC9 (PEQU_15883), PeUC10 (PEQU_33670), PeSCL1 (PEQU_04519), PeSCL2 (PEQU_06981), PeSCL3 (PEQU_12112), PeSCL4 (PEQU_19442), PeSCL5 (PEQU_39385), PePLCL1 (PEQU_01523), PePLCL1 (PEQU_07235), PePLCL3 (PEQU_15885), PePLCL4 (PEQU_40735), and PEQU_09781 from Phalaenopsis equestris. The five subclasses of PCs (ENODLs, UCLs, SCLs, PLCLs, and unknown), are indicated with different colors. Different clades are separated by the black lines. Bootstrap values are indicated only for those branches where they exceeded 50%. Scale bar represents 0.1 amino acid substitution per site.
had similar structures, and a majority of GPI-anchored proteins were localized to the plasma membrane (Chatterjee and Mayor 2001). In addition, 22 PePCs contained N-glycosylation sites (Table 1 Suppl.). As the presence of an N-terminal SP is a prerequisite for AGPs, we only examined PePCs possessing SPs. In total, 14 PePCs might represent chimeric AGPs, including seven PeENODLs, five PeUCLs, and two PeSCLs (Table 2 Suppl.). In addition, six PePCs may be both extensins and AGPs, one PePC might be an extensin or AGP, and seven PePCs contain putative extensin glycomodules.

Fig. 3. Heatmap showing the expression levels of PePC genes in four tissues (flowers, leaves, stems, and roots). The color scale represents RPKM-normalized log2 transformed counts. Blue represents low expression, and red indicates high expression. Gray indicates no expression in the tissue.

Based on the presence of an SP, PCLD, ALR, and GAS, the PePCs were divided into six groups (I - VI): 12 PePCs of the group I contain all four motifs (SP, PCLD, ALR, and GAS), 2 PePCs of the group II lack GAS, the single PePC of the group III lacks ALR, 6 PePCs of the group IV contain SP and PCLDs, 3 PePCs of the group V contain PCLD and GAS, and 6 PePCs of the group VI contain only PCLD (Table 2 Suppl., Fig. 1). Group I and II PePCs are considered to be chimeric AGPs due to the presence of SP and ALR, a half of the AGPs are PeENODLs and only the PePLCL subfamily is not represented (Table 2 Suppl.). Only the PeENODL subfamily consists of six groups, more than a half of which are AGPs. Among the six groups of putative PePCs, we found that all groups were included in the BrPC group, five groups were the same as the AtPC groups, and four groups resembled the structural types of OsPCs (Ma et al. 2011, Li et al. 2013).

We generated a phylogenetic tree based on the alignment of full-length PePC and AtPC protein sequences to analyze their evolutionary relationships (Fig. 2). Previously, 62 OsPCs, 84 BrPCs, as well as 38 AtPCs, were found to cluster into seven distinct clades (A - G) (Ma et al. 2011, Li et al. 2013). In the current study, the 30 PePCs also clustered into seven clades (A - G) with the 38 AtPCs (Fig. 2). In addition, most members of each subfamily clustered together with a few exceptions. All PCs in clades A, B, and C belonged to the ENODL
subfamily, most of which were ENODL-AGPs (except for AtENODL21 and PeENODL8), and almost all PeENODL-AGPs (except for PeENODL7) were grouped into clades A and B. Clade D only contained the SCL subfamily and three unknown AtPCs. Clade E contained two subfamilies (SCL and UCL) and one unknown PePC. Clades F and G both contained members of the UCL, PLCL, and ENODL subfamilies. More than a half of the members of clade F are ENODLs, this clade included most PePLCLs (except for PePLCL2). Clade G was almost entirely composed of UCLs (except for AtENODL18 and PePLCL2) (Table 2 Suppl., Fig. 2). Within each clade, particular clusters (bootstrap values > 50%) of orthologs (i.e., PeENODL8/AtENODL8 or PeSCL5/AtSC4) and paralogs (i.e., PeENODL4/PeENODL10 or PeUCL1/PeUCL2) were putatively identified. Therefore, the phylogenetic analysis indicated that each clade contained both orchid and Arabidopsis PCs and that members of the same subfamily tend to cluster in the same clade.

Exon-intron structural analysis of PC genes in *P. equestris* indicated that most PePC genes contained introns, except for PeSCL5 and PePLCL4 (Fig. 2 Suppl.). The number of exons in PePC ranged from one to three, with the majority (26 of 30) containing two exons, indicating that the number of exons in *PC* genes is proportional to the size of the genome. According to gene duplication analysis of these plants, segmental duplication and tandem duplication have contributed to the expansion of the *PC* gene family in Arabidopsis and rice, and segmental duplication events primarily led to the expansion of BrPC genes (Ma et al. 2011, Li et al. 2013, Cao et al. 2015). These findings suggest that gene duplication events, especially segmental duplication, have probably led to the different sizes of this gene family in different plant species.

We identified 30 PePCs, which were classified into four subfamilies, along with one unknown PC-like protein; these subfamilies were divided into six groups. Of the PCs in rice and Arabidopsis, the subfamilies and structures of PePCs are more like AtPCs than OsPCs. However, all PePCs contain only one PCLD and lack two PCLDs, whereas some AtPCs contain two PCLDs. In particular, the structures of PePCs group V members differ from both AtPCs and OsPCs, but they are similar to some BrPCs.

Phylogenetic analysis showed that each clade contains both *P. equestris* and Arabidopsis PCs, and subfamily members tend to cluster in the same clade. The orthologous genes between monocot and dicot *PC* genes indicate that each subfamily members shared a common ancestor before the divergence of monocot and dicot lineages, and the paralogous genes indicate that the expansion of the subfamilies occurred after the divergence of orchids and Arabidopsis, which is similar to the phylogenetic analysis results of the rice, Chinese cabbage, and Arabidopsis *PC* gene families (Ma et al. 2011, Li et al. 2013).

AGPs, extensins and Pro-rich proteins are known as hydroxyproline-rich glycoproteins (HRGPs). Much evidence indicates that HRGPs are involved in many aspects of cell growth and development, ranging from cell wall architecture and assembly to cell proliferation, cell-to-cell recognition, and cell expansion (Lamport 1965, Schultz et al. 1998, Majewska-Sawka and Nothnagel 2000, Showalter 2001). Over one third of PC proteins in *P. equestris* are chimeric AG glycoproteins, including seven ENODL, five UCL, and two SCL AGPs. By contrast, Arabidopsis contains 18 ENODL, seven UCL, and four SCL AGPs, and rice contains 18 ENODL, 19 UCL and one SCL AGP (Mashiguchi et al. 2009, Ma et al. 2011). Therefore, orchid contains fewer putative phytocyanin-like arabinogalactan proteins (PLAs) than Arabidopsis and rice. The protein backbones of classic AGPs contain amino acid residues Ala, Ser, Thr, Gly, and...
Hyp as their major constituents, as well as C-terminal hydrophobic regions instead of GASs, which allow AGP to localize to a lipid raft domain on the plasma membrane (Borner et al. 2003, Estévez et al. 2006). Due to the presence of SPs, ALRs, GASs, and PCLDs, 12 putative PLAs of group I in P. equestris were categorized as classic AGPs, and two putative PLAs of group II were classified as nonclassic AGPs. Meanwhile, some PePC-AGPs have putative extensin glycomodules, which function as sites for the addition of arabino-oligosaccharides. The functioning of these glycoproteins is affected by differences in their glycosylation states (Mashiguchi et al. 2009).

In the current study, we identified five genes that were highly expressed in flowers, stems, and roots, indicating that they may play important roles in P. equestris growth and development. The expression of genes in the ENODL and SCL subfamilies were relatively high, suggesting that these subfamilies may play specific roles in P. equestris. ENODL genes were firstly identified in legume nodules. Several ENODL genes are activated in response to inoculation with arbuscular mycorrhizal fungi (Harrison 1998), and some may play important roles in the development of symbiotic orchid mycorrhizae (Perotto et al. 2014). However, we found that ENODL genes were expressed at very low levels in roots, perhaps because the orchids used in this study were not specifically inoculated with fungi, which would have induced ENODL gene expression. In order to explore the mutualistic plant-microbe interactions in orchids more thoroughly, further efforts are needed to verify the expression of these genes and to elucidate the molecular mechanism of orchid mycorrhizal formation.

References


Garrett, T.P.J., Clingeleffer, D.J., Guss, J.M., Rogers, S.J., Freeman, H.C.: The crystal structure of poplar apoplastic cyanin at 1.8 Å resolution. The geometry of the copper-binding site is created by the polypeptide. - J. Biol. Chem. 259: 1822-1825, 1984.


Majewska-Sawka, A., Nothnagel, E.A.: The multiple roles of...


