Arch. Biol. Sci., Belgrade, 67(2), 373-383, 2015 DOI:10.2298/ABS141111005W

MOLECULAR CLONING OF *RBCS* GENES IN *SELAGINELLA* AND THE EVOLUTION OF THE *RBCS* GENE FAMILY

Bo Wang[1,2], Yingjuan Su[3,4,*], and Ting Wang[1,*]

[1] *CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences*, Wuhan, China

[2] *University of Chinese Academy of Sciences,* Beijing, China

[3] *State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University,* Guangzhou, China

[4] *Institute for Technology Research and Innovation of Sun Yat-sen University*, Zhuhai, China

*__Corresponding authors:__ suyj@mail.sysu.edu.cn; tingwang@wbgcas.cn

Abstract: Rubisco small subunits (RBCS) are encoded by a nuclear *rbcS* multigene family in higher plants and green algae. However, owing to the lack of *rbcS* sequences in lycophytes, the characteristics of *rbcS* genes in lycophytes is unclear. Recently, the complete genome sequence of the lycophyte *Selaginella moellendorffii* provided the first insight into the *rbcS* gene family in lycophytes. To understand further the characteristics of *rbcS* genes in other *Selaginella*, the full length of *rbcS* genes (*rbcS1* and *rbcS2*) from two other *Selaginella* species were isolated. Both *rbcS1* and *rbcS2* genes shared more than 97% identity among three *Selaginella* species. RBCS proteins from *Selaginella* contained the Pfam RBCS domain F00101, which was a major domain of other plant RBCS proteins. To explore the evolution of the *rbcS* gene family across *Selaginella* and other plants, we identified and performed comparative analysis of the *rbcS* gene family among 16 model plants based on a genome-wide analysis. The results showed that (i) two *rbcS* genes were obtained in *Selaginella*, which is the second fewest number of *rbcS* genes among the 16 representative plants; (ii) an expansion of *rbcS* genes occurred in the moss *Physcomitrella patens*; (iii) only RBCS proteins from angiosperms contained the Pfam PF12338 domains, and (iv) a pattern of concerted evolution existed in the *rbcS* gene family. Our study provides new insights into the evolution of the *rbcS* gene family in *Selaginella* and other plants.

Keywords: *Selaginella*; *rbcS* gene; sequence analysis; phylogenetic analysis; genome-wide analysis

Received November 11, 2014; **Revised** December 8, 2014; **Accepted** December 9, 2014

INTRODUCTION

Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco; EC 4.1.1.39) is a stromal protein that catalyzes two competing reactions, photosynthetic CO_2 fixation and photorespiratory carbon oxidation (Andersson and Backlund, 2008). Since Rubisco has a relatively lower turnover rate and there is always competition between O_2 and CO_2 at the active site, it is the rate-limiting enzyme of photosynthesis (Andersson and Backlund, 2008). Therefore, Rubisco is often viewed as a potential target for genetic manipulation to improve crop yields (Mann, 1999).

In higher plants and green algae, the Rubisco holoenzyme is composed of large (RBCL) and small (RBCS) subunits encoded respectively by the unique chloroplastic *rbcL* gene and the nuclear *rbcS* multigene family (Dean et al., 1989). Although small subunits do not contain the active sites for catalytic activity, they play an important role in holoenzyme assembly, stability, and influence Rubisco catalytic efficiency and specificity (Spreitzer, 2003). The number of *rbcS* gene family members ranges from two genes in *Chlamydomonas* to twenty-two or more in wheat (Spreitzer, 2003). These copies are distributed at one or more loci, and individual gene copies are generally arranged in a tandem array at a locus. Previous studies have shown that members of the *rbcS* gene family in one plant are generally more similar to each other than to members of a family in a different species. In other words, the physically adjacent *rbcS* genes within a species show the highest similarity, followed by *rbcS* genes at different loci within a species, and then by *rbcS* genes between species, which are the most diverged (Clegg et al., 1997).

Lycophytes belong to an ancient lineage of vascular plants that diverged from the seed plant lineage immediately after plants colonized terrestrial environments (Banks, 2009). However, little is known of the characteristics of *rbcS* genes in lycophytes due to lacking data of *rbcS* sequences. Recently, the genome of *Selaginella moellendorffii* has been sequenced (Banks et al., 2011). It offers an opportunity to extend our understanding of the *rbcS* gene family into lycophytes. With *S. moellendorffii* genome sequences, *rbcS* genes have been identified from two other *Selaginella* species by PCR approaches. The sequence features of *rbcS* genes, the phylogenetic relationship, and comparison with a phylogenetic tree of the chloroplastic *rbcL* gene are also discussed in this study. To date, more and more plant genomes have been fully sequenced, and their genomic sequences and annotation are publicly available. These facilitate comparative genomic studies of plants, making it possible to address major plant biology questions *in silico*. Thus, we have performed comparative analysis of the *rbcS* multigene family in 16 sequenced plant and algal genomes, to define the number of *rbcS* gene homologs across these genomes, and to investigate the evolution of the *rbcS* gene family. The conserved domains and the phylogeny of the *rbcS* gene family across different genomes are also studied.

MATERIALS AND METHODS

Plant materials and genomic DNA isolation

Young leaves of *Selaginella doederleinii* and *S. involvens* were collected from Wuhan Botanical Garden, Chinese Academy of Sciences. Total DNA was isolated using the DNA Extraction Kit (Tiangen).

Cloning and sequencing of full-length rbcS genes from Selaginella

The *rbcS* gene of the fern *Pteris vittata* was exploited as a query to identify *rbcS* genes in the *S. moellendorffii* genome using BLAST search. A series of primers was designed based on upstream and downstream sequences of *rbcS* genes in *S. moellendorffii*. The *rbcS* genes of two other *Selaginella* were amplified by PCR. All primers used are listed in Table 1.

Table 1. Primers used in PCR for rbcS1 and rbcS2 genes amplification from two Selag inella species.

Gene name	Primer name	Primer sequence(5'-3')
rbcS1	Upstream-*rbcS1*	AGGCCGAGCTCACCATCACC
	Downstream-*rbcS1*	GAGCCTCGTATGCCATTATCGT
rbcS2	Upstream-*rbcS2*	ATTCTGAGCCCAAGACCTA
	Downstream-*rbcS2*	GRCGGAGCTCAATTGGTAGAG

The PCR reaction was conducted in a total volume of 50 μL mixture containing 31.5 μL of dd H_2O, 8.0 μL of 2.5 mM dNTP mixture, 5.0 μL of Ex Taq buffer (Takara), 2.5 U Ex Taq (Takara), 0.4 μM of each primer, and 1.0 μL of template DNA (50 ng/μL). A typical PCR amplification included an initial denaturation (5 min, 94°C) followed by 35 cycles with a 1 min denature at 94°C, 1 min annealing at 55°C, 2.5 min synthesis at 72°C, and a final synthesis step for 10 min at 72°C. Products were visualized on an ethidium bromide-stained agarose gel. Three amplified products were recovered using the DNA rapid purification kit (Axygen). The purified PCR products were ligated into PCR2.1 vector (Invitrogen) and then used to transform competent *E. coli* cells DH-5α for sequencing.

Sequence analysis

The BioEdit software was used to analyze the DNA and protein sequences, including GC contents and amino acid composition. Based on the annotation of *SmrbcS1* and *SmrbcS2* in GenBank, the prediction of the protein was analyzed by the Fgenesh gene-finder (Solovyev et al., 2006). Multiple protein sequence alignments were calculated by Muscle, and an alignment plot was created by Espript 3.0 (Robert and Gouet, 2014).

Phylogenetic analysis

The sequences of *rbcS1* and *rbcS2* genes from *S. doederleinii* and *S. involvens* were submitted to GenBank (accession number KM396421 to KM396424). In addition to sequences from *Selaginella*, sequences used for comparison and phylogenetic analysis were downloaded from GenBank. The species names and accession numbers are listed in Table 2. The LG + I + G model was selected as the best substitution model for RBCS protein phylogenetic analysis by ProtTest 2.4 (Abascal et al., 2005). The RBCS phylogenetic tree was constructed by PhyML 3.1 (Guindon et al., 2010) under the maximum likelihood method with 100 bootstraps. In order to make a comparison with an RBCS phylogenetic tree, we also constructed the phylogenetic tree based on the chloroplastic *rbcL* gene.

Analysis of the evolution of rbcS gene family in model plants based on genome-wide analysis

Data sources

RBCS protein sequences were retrieved from published studies and publicly available databases. Sequences of *Arabidopsis thaliana* RBCS proteins were obtained from the *Arabidopsis* Information Resource (http://www.arabidopsis.org). Rice RBCS proteins were retrieved from the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu). The proteome sequences of *Manihot esculenta, Populus trichocarpa, Glycine max, Malus domestica, Arabidopsis lyrata, Carica papaya, Vitis vinifera, Sorghum bicolor, Zea mays, Brachypodium distachyon, Selaginella moellendorffii, Physcomitrella patens, Chlamydomonas reinhardtii* v5.3 and *Volvox carteri* were downloaded from Phytozome (http://www.phytozome.net). Proteome data of several plants include alternatively spliced variants. For instance, *Arabidopsis thaliana* RBCS3B has three known splicing variants (AT5G38410.1, AT5G38410.2 and AT5G38410.3) and AT5G38410.1 is the primary transcript. We counted them as one single gene, and only the primary transcript were included in our phylogenetic analyses.

BLAST search

We downloaded known RBCS protein sequences from UniProt (Consortium, 2012). We took this dataset as the initial query to search against the proteome sequences of these plant genomes.

Table 2. Species and accession numbers of amino acid sequences used for phylogenetic analysis.

lineage	Species	Protein Names	Accession number	
			RBCS proteins	rbcL genes
Algae	*Chlamydomonas reinhardtii*	CrRBCS1	P00873	NC_005353
	Chlamydomonas reinhardtii	CrRBCS2	P08475	
Liverworts	*Marchantia paleacea*	MpRBCS	O64416	DQ286015
Mosses	*Physcomitrella patens*	PpRBCS	BAC87878	AP005672
Ferns	*Pteris vittata*	PvRBCS	CAA67061	EF473709
Gymnosperms	*Larix laricina*	LlRBCS	P16031	AF479878
	Pinus thunbergii	PtRBCS	P10053	NC_001631
Monocots	*Fritillaria agrestis*	FaRBCS1/4	O24634	AF013233
	Fritillaria agrestis	FaRBCS2	O22572	
	Fritillaria agrestis	FaRBCS3	O22573	
	Fritillaria agrestis	FaRBCS5	O22645	
	Oryza sativa Japonica Group	OsRBCSA	P18566	NC_001320
	Oryza sativa Japonica Group	OsRBCSC	Q0INY7	
	Zea mays	ZmRBCS	P05348	NC_001666
Dicots	*Amaranthus hypochondriacus*	AhRBCS1	Q42516	X51964
	Amaranthus hypochondriacus	AhRBCS2	Q9XGX5	
	Amaranthus hypochondriacus	AhRBCS3	Q9XGX4	
	Arabidopsis thaliana	AtRBCS1A	P10795	NC_000932
	Arabidopsis thaliana	AtRBCS1B	P10796	
	Arabidopsis thaliana	AtRBCS2B	P10797	
	Arabidopsis thaliana	AtRBCS3B	P10798	
	Flaveria pringlei	FpRBCS1	Q39743	HQ534133
	Flaveria pringlei	FpRBCS2	Q39744	
	Flaveria pringlei	FpRBCS3	Q39745	
	Flaveria pringlei	FpRBCS4	Q39746	
	Flaveria pringlei	FpRBCS5	Q39747	
	Flaveria pringlei	FpRBCS6	Q39748	
	Flaveria pringlei	FpRBCS7	Q39749	
	Mesembryanthemum crystallinum	McRBCS1	P16032	HM850175
	Mesembryanthemum crystallinum	McRBCS2	Q04450	
	Mesembryanthemum crystallinum	McRBCS3	Q08183	
	Mesembryanthemum crystallinum	McRBCS4	Q08184	
	Mesembryanthemum crystallinum	McRBCS5	Q08185	
	Mesembryanthemum crystallinum	McRBCS6	Q08186	
	Solanum lycopersicum	SlRBCS1	P08706	NC_008096
	Solanum lycopersicum	SlRBCS2A	P07179	
	Solanum lycopersicum	SlRBCS3A/3C	P07180	
	Solanum lycopersicum	SlRBCS3B	P05349	

HMMER search

HMMER search was widely applied for identification of homologs of the protein family of interest (Eddy, 2009). There are two Pfam (Punta et al., 2012) domain models for the RBCS proteins, PF12338 (RbcS) and PF00101 (RuBisCO_small), both of which were searched in our analyses. We performed HMMER search (Eddy, 1998) using the Pfam profile PF12338 and PF00101 against the annotated protein sequences of the 16 genomes and refined the results manually to obtain RBCS proteins.

Intersection of search results

We used E-value 1.0 as the cutoff in BLASTP and HMMER search searches, and kept only the hits returned by both searches. As a result, the final hits should be similar to the query *rbcS* genes in the pairwise sequence comparison and contain either of the two conserved Pfam domains. Through this method, all known RBCS proteins were successfully retrieved and no false positives were found.

Phylogenetic analysis

Multiple sequence alignments (MSAs) of the full-length protein sequences were performed by the MAFFT program (Katoh et al., 2005) using two highly accurate methods: L-INS-i and E-INS-i. The Jones-Taylor-Thornton (JTT) model was selected as the best-fitting amino acid substitution model by ProtTest 2.4 (Abascal et al., 2005). The maximum likelihood (ML) phylogenetic tree was constructed using PhyML 3.1 (Guindon et al., 2010) under the JTT model with 100 replicates of bootstrap analysis, estimated proportion of invariable sites, four rate categories, estimated gamma distribution parameter, and optimized starting BIONJ tree (Gascuel, 1997). The phylogenetic tree was visualized using the program Figtree (http://tree.bio.ed.ac.uk/software/ figtree).

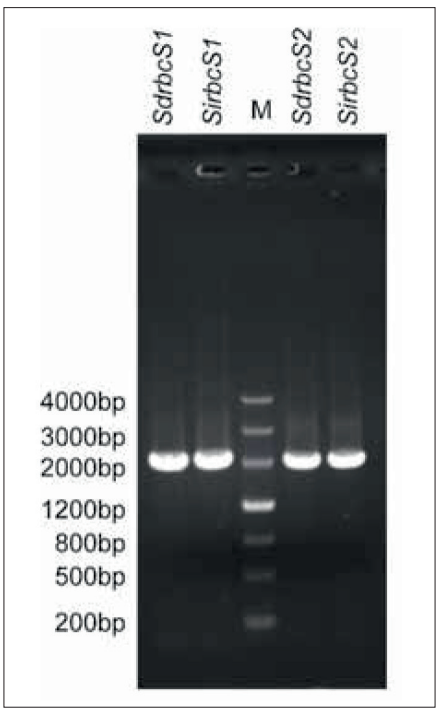

Fig. 1. Agarose gel electrophoresis (1.0%) of the PCR products of *rbcS1* and *rbcS2* from two *Selaginella* species.

RESULTS AND DISCUSSION

Cloning and characterization of full length of rbcS genes from Selaginella

To obtain *rbcS* sequences of *S. moellendorffii*, we performed a BLAST search using the *rbcS* gene of the fern *Pteris vittata* against *S. moellendorffii* genome sequences. Two *rbcS* genes were identified in the *S. moellendorffii* genome, named as *SmrbcS1* and *SmrbcS2*. In order to understand the characteristics of *rbcS* genes in other *Selaginella*, we isolated *rbcS* gene sequences from two other *Selaginella*: *S. doederleinii* and *S. involvens*. Two pairs of the primers for PCR were designed based on *SmrbcS1* and *SmrbcS2* sequences. The *rbcS1* and *rbcS2* genes from *Selaginella* species were amplified by PCR and *rbcS1* and *rbcS2* fragments were obtained in *S. doederleinii* and *S. involvens*,

respectively (Fig. 1), covering the full lengths of *rbcS1* and *rbcS2* genes, when compared to corresponding regions from *SmrbcS1* and *SmrbcS2*.

Nucleotide sequence analyses indicated that their guanine-cytosine (GC) contents were 52.55-52.90% for *rbcS1* and 48.25-48.69% for *rbcS2*, respectively (Table 3). The *rbcS1* and *rbcS2* genes from three *Selaginella* species shared more than 98% and 97% identity, respectively (Table 3).

Amino acid analysis of RBCS proteins in Selaginella

Using *SmrbcS1* and *SmrbcS2* as a reference, the nucleotides of *rbcS* genes from other *Selaginella* were analyzed by the Fgenesh gene-finder to deduce the amino acid sequences of proteins. The lengths of predicted RBCS proteins were 160-178 amino acids (Table 3). With *S. moellendorffii* as a reference, both RBCS1 and RBCS2 proteins showed above 97% identity among three *Selaginella* species. Taking *S. moellendorffii* as an example, SdRBCS1 and SdRBCS2 proteins comprised 178 and 160 amino acids, respectively. The amino acid composition analysis showed that Ala was the most abundant amino acid residue in both SdRBCS1 and SdRBCS2, with frequencies of 5.00% and 7.87%, respectively.

Sequences comparison with other plant RBCS proteins

Fig. 2 shows an alignment of amino acid sequences including six RBCS sequences from *Selaginella*, one from *Marchantia paleacea*, one from *Physcomitrella patens*, one from *Pteris vittata*, one from *Larix laricina*, and four from *Arabidopsis thaliana*. RBCS1 proteins showed more similarity among themselves than RBCS2 proteins. The polypeptide chain of RBCS was folded into six major domains, α-helix A, B, and β-strands A to D (Spreitzer, 2003) (Fig. 2). These domains all existed in each of the six RBCS proteins of *Selaginella*, and they showed higher sequence similarity among plants.

Phylogenetic relationships of RBCS proteins and comparison with the rbcL phylogenetic tree

RBCS protein sequences have been reported from divergent species across the plant kingdom. To clarify the evolutionary relationships of the *rbcS* gene family in plants, we performed a phylogenetic analysis using the RBCS protein sequences of three *Selaginella* species along with another 38 RBCS protein sequences, representing RBCS proteins from 14

Table 3. Characteristics of *rbcS* genes in three *Selaginella* species.

Gene name	Species	Nucleotide			Protein	
		Length (bp)	GC content (%)	Similar* (%)	Length (bp)	Similar* (%)
SmrbcS1	*S. moellendorffii*	1429	52.76	-	178	-
SdrbcS1	*S. doederleinii*	1429	52.55	99.0	178	99.4
SirbcS1	*S. involvens*	1429	52.90	98.8	177	98.9
SmrbcS2	*S. moellendorffii*	1482	48.25	-	160	-
SdrbcS2	*S. doederleinii*	1482	48.38	97.8	160	97.5
SirbcS2	*S. involvens*	1483	48.69	98.1	160	98.1

*With *S. moellendorffii* as the reference, sequence identity was estimated by use of BioEdit.

diverse species including the angiosperms, gymnosperms, ferns, mosses, liverworts and algae (Fig. 3).

Although several bootstrap values were low, a distinct evolution of RBCS along with the hierarchy of plant taxon from algae to higher plants was apparent (Fig. 3A). Within land plants, gymnosperms and angiosperms were monophyletic with high bootstrap support. Gymnosperms, which include *Larix laricina* and *Pinus thunbergii*, were a sister group of angiosperms. Among angiosperms, monocot and dicot failed to form two distinct groups. The monocot *Fritillaria agrestis* was the basal clade in angiosperms, while another two monocots, *Oryza sativa* and *Zea mays*, were grouped together with the dicot *Solanum lycopersicum*. In angiosperms, the RBCS proteins formed species-specific paralogous clusters, indicating that *rbcS* genes had extensively evolved since these species diverged. Spreitzer (2003) also reported that all members of a plant *rbcS* gene family were generally more similar to each other than to members of a family in a different species, resulting in only a few amino acid differences in the small subunits within a family; various mechanisms of gene conversion that maintain the genes nearly identical have been proposed (Clegg et al., 1997; Heinhorst et al., 2002). The ferns, lycophytes, mosses, liverworts and algae constituted the basal branch

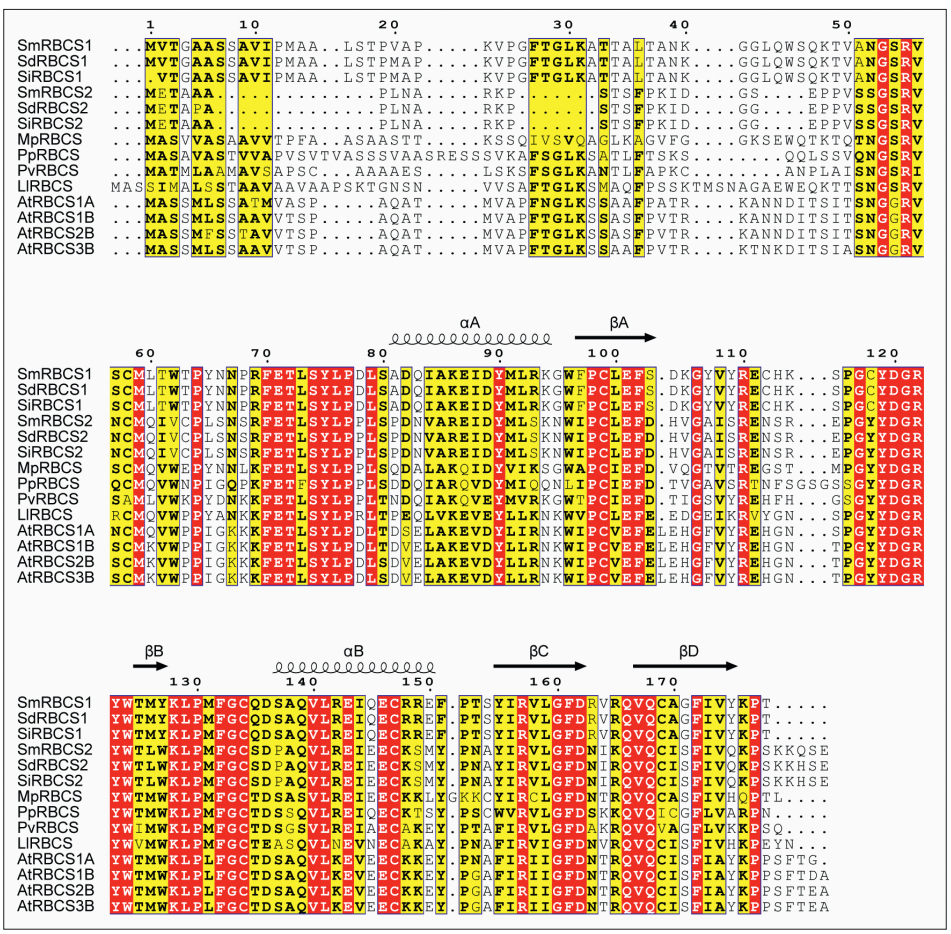

Fig. 2. Alignment of amino acid sequences of plant RBCS proteins. Conserved residues across all sequences are in color. The protein secondary structure α-helix A, B, and β-strands A to D are also indicated by horizontal arrows.

of the phylogenetic tree, but their branching order failed to correspond to the organismal phylogeny (Fig. 3A).

In accordance with the sequence alignment, RBCS1 and RBCS2 formed distinct groups in *Selaginella*. Three *Selaginella* RBCS1 were clustered together and three *Selaginella* RBCS2 were clustered. Both *rbcS* genes formed orthologous clusters, suggesting that *rbcS1* and *rbcS2* genes were present before these *Selaginella* species diverged. Although three *Selaginella* RBCS2 were grouped together with the moss *Marchantia paleacea*, the bootstrap value was quite low.

Besides the phylogenetic tree of *rbcS* genes, we also constructed a phylogenetic tree based on the multiple sequence alignments of the chloroplastic *rbcL* gene (Fig. 3B). When comparing *rbcS* and *rbcL* phylogenetic trees, the topologies were largely congruent. As in the *rbcS* phylogenetic tree, the basal branch of the *rbcL* phylogenetic tree, from algae to ferns, was still inconsistent with the organismal phylogeny. Among the clade of seed plants, there were two major differences between *rbcS* and *rbcL* phylogenetic trees. First, the gymnosperms *Larix laricina* and *Pinus thunbergii* were clustered together in the RBCS phylogenetic tree, while *Pinus thunbergii* was grouped with *Physcomitrella patens* in the *rbcL* phylogenetic tree. Second, the other major difference was the relationship between monocots and dicots. Dicots were the sister to the clade of monocots in the *rbcL* phylogenetic tree, while they did not completely separate in the *rbcS* phylogenetic tree.

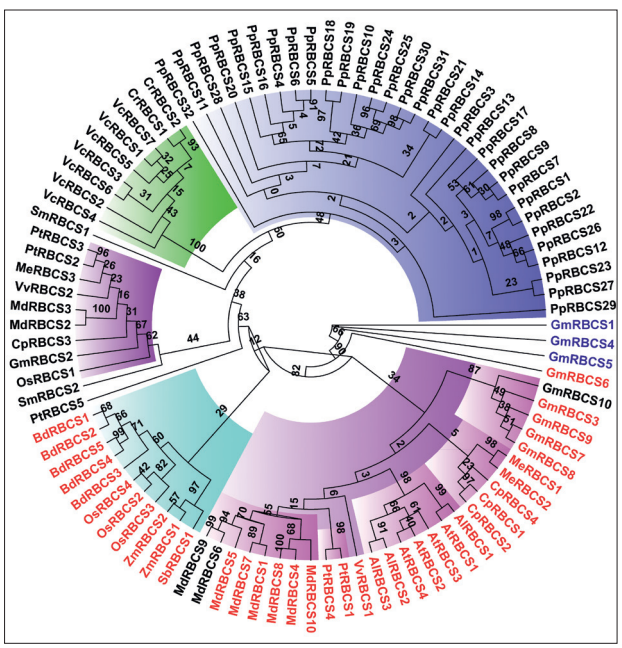

Fig. 3. Phylogenetic trees. A – established from amino acid sequences of plant RBCS proteins; B – inferred from the nucleotide sequences of plant *rbcL* genes. Numbers near to the nodes indicate bootstrap values.

Evolution of rbcS gene family in model plants based on genome-wide analysis

Plants and green algae have multiple nuclear-encoded *rbcS* genes, ranging from as many as 22 or more in wheat to as few as 2 in the green alga *Chlamydomonas reinhardtii* (Dean et al., 1989; Spreitzer, 2003). In order to identify all putative RBCS proteins in model plant genomes and explore the evolution of the *rbcS* gene family, we performed BLASTP and HMMER search against the complete genome or genome assemblies of the eudicots *Manihot esculenta*, *Populus trichocarpa*, *Glycine max*, *Malus domestica*, *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Carica papaya* and *Vitis vinifera*, the monocots *Sorghum bicolor*, *Zea mays, Oryza sativa* and *Brachypodium distachyon*, the lycophyte *Selaginella moellendorffii*, the moss *Physcomitrella patens*, and the algae *Chlamydomonas reinhardtii* and *Volvox carteri*, respectively. BLASTP and HMMER search results showed that 96 RBCS proteins existed in various green plants from unicellular green algae to angiosperms (Table 4).

We identified two and seven RBCS proteins in algae *Chlamydomonas reinhardtii* and *Volvox carteri*, respectively. Our data showed that the

Table 4. The number of *rbcS* genes in 16 representative plants.

Lineage	Organism	Number of rbcS genes				Nomenclature
		Total	PF12338 and PF00101[a]	PF12338[b]	PF00101[c]	
Dicots	*Manihot esculenta*	3	2	0	1	MeRBCS
Dicots	*Populus trichocarpa*	5	2	0	3	PtRBCS
Dicots	*Glycine max*	10	5	3	2	GmRBCS
Dicots	*Malus domestica*	10	6	0	4	MdRBCS
Dicots	*Arabidopsis lyrata*	3	3	0	0	AlRBCS
Dicots	*Arabidopsis thaliana*	4	4	0	0	AtRBCS
Dicots	*Carica papaya*	4	3	0	1	CpRBCS
Dicots	*Vitis vinifera*	2	1	0	1	VvRBCS
Monocots	*Sorghum bicolor*	1	1	0	0	SbRBCS
Monocots	*Zea mays*	2	2	0	0	ZmRBCS
Monocots	*Oryza sativa*	4	3	0	1	OsRBCS
Monocots	*Brachypodium distachyon*	5	5	0	0	BdRBCS
Lycophytes	*Selaginella moellendorffii*	2	0	0	2	SmRBCS
Mosses	*Physcomitrella patens*	32	0	0	32	PpRBCS
Algae	*Chlamydomonas reinhardtii*	2	0	0	2	CrRBCS
Algae	*Volvox carteri*	7	0	0	7	VcRBCS
Total	All plants	96	37	3	56	

a: The number of RBCS proteins contained the Pfam PF12338 domain and the Pfam PF00101 domain.
b: The number of RBCS proteins only include the Pfam PF12338 domain.
c: The number of RBCS proteins only include the Pfam PF00101 domain.

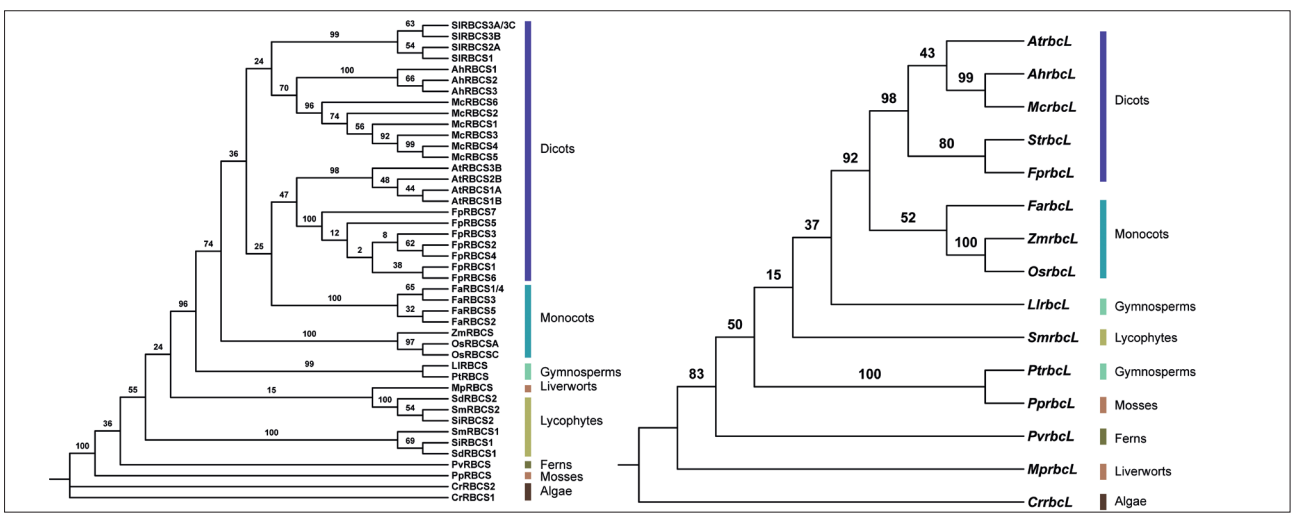

Fig. 4. Maximum likelihood analysis of 96 RBCS protein sequences presented as an unrooted phylogenetic tree. The taxa with red color are RBCS proteins containing both the Pfam PF12338 domain and PF00101 domains. The taxa in blue and black are RBCS proteins only including the PF12338 and PF00101 domains, respectively.

genomes of vascular plants generally encoded no more than five RBCS proteins, except for *Glycine max* and *Malus domestica* that both encoded ten RBCS proteins. In contrast, genomes of moss, namely *Physcomitrella patens*, encoded as many as 32 RBCS proteins. This suggested an expansion of RBCS proteins occurred in the bryophyte (Table 4). Two RBCS proteins were obtained in *Selaginella moellendorffii,* which had the second fewest number of *rbcS* genes among the 16 representative plants.

To obtain protein domain information, we searched these RBCS proteins from model plants against the Pfam database. The results showed that there were two kinds of protein domains for all RBCS proteins, the Pfam PF00101 domain (99 aa) and the Pfam PF12338 domain (45 aa). We found that only 37 of the 96 predicted proteins contain both the Pfam PF12338 domain and the PF00101 domain, but 3 and 56 RBCS proteins only include the PF12338 domain and the PF00101 domain, respectively (Table 4). As only angiosperms contained the domain (PF12338) (Table 4, Fig. 4), we concluded that the PF12338 domain of *rbcS* genes appeared in the seed plant lineage after the divergence of lycophytes but before the divergence of monocots and dicots.

We also searched the six RBCS proteins from three *Selaginella* in this study against the Pfam database. The results indicated that all RBCS proteins from three *Selaginella* contained the Pfam F00101 domain, which was a typical protein domain for Rubisco small subunits. We constructed a phylogenetic tree using the RBCS proteins of 16 representative plants to unveil the evolutionary relationships among plant RBCS proteins (Fig. 4). The tree topology and the corresponding phylogenetic relationships also indicated that proteins from the same species clustered together with high support values. This indicated that concerted evolution existed in the *rbcS* gene family. Previous studies have shown a pattern of concerted evolution among the three *rbcS genes* in Solanaceae, where paralogs (genes related by duplication) are more similar than orthologs (genes relate d by speciation). Concerted evolution is the non-independent evolution of gene copies in a multigene family that can homogenize members and prevent the acquisition of new functions.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (Nos. 31000171, 31070594, and 31370364), the Knowledge Innovation Program of the Chinese Academy of Sciences (Nos. KSCX2-EW-J-20 and KSCX2-YW-Z-0940), the Basic Research Project of the Department of Science and Technology of Zhuhai city, China (No. 2012D0401990031), and the the Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences. We thank Lei Gao, Yuan Zhou, and Zhiwei Wang for their help in drafting the manuscript.

Authors' contribution: BW participated in the design of the study, performed the experiments and data analysis, and drafted the manuscript. YJS and TW conceived the study, participated in its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Conflict of interest disclosure: The authors declare no conflict of interest.

REFERENCES

Abascal, F., Zardoya, R. and *D. Posada* (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. **21**, 2104-2105.

Andersson, I. and *A. Backlund* (2008). Structure and function of Rubisco. *Plant Physiol. Biochem*. **46**, 275-291.

Banks, J.A. (2009). *Selaginella* and 400 million years of separation. *Annu. Rev. Plant Biol*. **60**, 223-238.

Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., DePamphilis, C., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A., Ashton, N.W., Axtell, M.J., Barker, E., Barker, M.S., Bennetzen, J.L., Bonawitz, N.D., Chapple, C., Cheng, C.Y., Correa, L., Dacre, M., DeBarry, J., Dreyer, I., Elias, M., Engstrom, E.M., Estelle, M., Feng, L., Finet, C., Floyd, S.K., Frommer, W.B., Fujita, T., Gramzow, L., Gutensohn, M., Harholt, J., Hattori, M., Heyl, A., Hirai, T., Hiwatashi, Y., Ishikawa, M., Iwata, M., Karol, K.G., Koehler, B., Kolukisaoglu, U., Kubo, M., Kurata, T., Lalonde, S., Li, K.J., Li, Y., Litt, A., Lyons, E., Manning, G., Maruyama, T., Michael, T.P., Mikami, K., Miyazaki, S., Morinaga, S., Murata, T., Mueller-Roeber, B., Nelson, D.R., Obara, M., Oguri, Y., Olm stead, R.G., Onodera, N., Petersen, B.L., Pils, B., Prigge, M.,

Rensing, S.A., Riano-Pachon, D.M., Roberts, A.W., Sato, Y., Scheller, H.V., Schulz, B., Schulz, C., Shakirov, E.V., Shibagaki, N., Shinohara, N., Shippen, D.E., Sorensen, I., Sotooka, R., Sugimoto, N., Sugita, M., Sumikawa, N., Tanurdzic, M., Theissen, G., Ulvskov, P., Wakazuki, S., Weng, J.K., Willats, W., Wipf, D., Wolf, P.G., Yang, L.X., Zimmer, A.D., Zhu, Q.H., Mitros, T., Hellsten, U., Loque, D., Otillar, R., Salamov, A., Schmutz, J., Shapiro, H., Lindquist, E., Lucas, S., Rokhsar, D. and *I.V. Grigoriev* (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*. **332**, 960-963.

Clegg, M.T., Cummings, M.P. and *M.L. Durbin* (1997). The evolution of plant nuclear genes. *Proc Natl Acad Sci U S A*. **94**, 7791-7798.

Consortium, T.U. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* **40**, D71-D75.

Dean, C., Pichersky, E. and *Dunsmuir, P. (1989).* Structure, evolution, and regulation of RbcS genes in higher plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **40**, 415-439.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics.* **14**, 755-763.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**,205-211.

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685-695.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and *O. Gascuel* (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321.

Heinhorst, S., Baker, S.H., Johnson, D.R., Davies, P.S., Cannon, G.C. and *J.M. Shively* (2002). Two copies of form I RuBisCO genes in *Acidithiobacillus ferrooxidans* ATCC 23270. *Curr. Microbiol.* **45**, 115-117.

Katoh, K., Kuma, K., Toh, H. and *T. Miyata* (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.* **33**, 511-518.

Mann, C.C. (1999). Future food - Bioengineering - Genetic engineers aim to soup up crop photosynthesis. *Science.* **283**, 314-316.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E., Eddy, S.R., Bateman, A. and *R.D. Finn* (2012). The Pfam protein families database. *Nucl. Acids Res.* **40**, D290-D301.

Robert, X. and *P. Gouet* (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucl. Acids Res.* ***42***, W320-324.

Solovyev, V., Kosarev, P., Seledsov, I. and *D. Vorobyev* (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, S10.

Spreitzer, R.J. (2003). Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Arch. Biochem. Biophys.* **414**, 141-149.