Molecular evolution of GDP-L-galactose phosphorylase (GGP), a key regulatory gene in plant ascorbate biosynthesis

Junjie Tao (taojj@jxau.edu.cn)
Jiangxi Agricultural University https://orcid.org/0000-0002-1002-6496
Zhuan Hao
Weinan Normal University
Chunhui Huang
Jiangxi Agricultural University

Research article

Keywords: AsA, Ascorbate, VTC2, GGP, L-galactose pathway, Molecular evolution, Gene duplication

Posted Date: October 28th, 2019

DOI: https://doi.org/10.21203/rs.2.16500/v1

Version of Record: A version of this preprint was published at AoB PLANTS on November 4th, 2020. See the published version at https://doi.org/10.1093/aobpla/plaa055.

Abstract

Background Ascorbic acid (AsA) is a multi-functional molecule and plays essential roles in maintaining the normal life activities of living organisms. Although widely present in plants, the concentration of AsA varies greatly in different plant species. The GDP-L-galactose phosphorylase (GGP) is a key regulatory gene in plant AsA biosynthesis that can regulate the concentration of AsA at the transcriptional and translational levels. The function and regulation mechanisms of GGP have been well understood in previous works. However, the molecular evolutionary patterns of the gene remain unclear.

Results In this study, a total of 149 homologous sequences of GGP were sampled from 71 plant species covering the major groups of Viridiplantae, including angiosperms, gymnosperms, lycophytes, bryophytes and chlorophytes, and their phylogenetic relationships, gene duplication and molecular evolution analyses were investigated. Phylogenetic analysis showed that GGP exists widely in various plants, and five major duplication events and several taxon-specific duplications were found, which led to the rapid expansion of GGP genes in seed plants, especially in angiosperms. The structure of GGP genes were more conserved in land plants, but varied greatly in green algae, indicating that GGP may have undergone great differentiation in the early stages of plant evolution. Most GGP proteins have a conserved motif arrangement and composition, suggesting that plant GGPs have similar catalytic functions. Molecular evolutionary analyses showed that plant GGP genes was predominated by strong purifying selection, indicating the functional importance and conservativeness of plant GGP genes during evolution. Most of the branches under positive selection identified by branch-site model were mainly in the chlorophytes lineage, indicating the evolutionary innovation of GGP genes also mainly occurred in the early stages of plant evolution and episodic diversifying selection contributed to the evolution of plant GGP genes.

Conclusions The molecular evolutionary patterns of GGP were first systematically explored in this study. The conservative function of GGP and its rapid expansion in angiosperms may be one of the reasons for the increase of AsA content in angiosperms, enabling angiosperms to adapt to changing environments.

Background

L-Ascorbic acid (AsA), also well-known as ascorbate or Vitamine C (Vc), is a water soluble vitamin and an essential micronutrient for the normal growth and development of both animals and plants. As a major antioxidant, AsA can protect cells in living organisms from the threat of reactive oxygen species (ROS) under abiotic stress. At the same time, AsA is also a cofactor for dioxygenase and plays a vital role in most metabolic processes [1]. AsA presents widely in plant tissues and is a multi-functional metabolite, which linked to many physiological processes, such as regulating photosynthesis, growth and development, cell wall biosynthesis, regulating seed germination, flowering time, fruit softening and aging, postharvest storage, mediating signal transduction, and enhancing plant resistance to adverse environments [2–4]. Lack of AsA in human body can lead to scurvy and other diseases, while an appropriate amount of AsA is beneficial to aging, cancer and other diseases [5]. However, humans and some mammals have lost the ability to synthesize AsA by themselves due to several severe mutations

occurred in the gene encoding L-gulonolactone oxidase (GuLO) in AsA synthesis reaction [6], and thus have to secure the required AsA from plant sources, especially fresh fruits and vegetables which are rich in AsA, to cover their daily requirements. In view of the unique functions and importance of AsA in normal life activities of plants and animals, it is of great interest to study the biosynthesis and regulation of AsA in plants.

Four possible biosynthetic pathways to AsA have been proposed in plants, named the L-galactose pathway [7] and the other three alternative pathways including L-glucose pathway [8], the D-galacturonic acid [9] and the myo-inositol pathway [10]. The L-galactose pathway, also named as the Smirnoff-Wheeler pathway, is the best established AsA biosynthesis pathway in plants and considered to be the only significant pathway for AsA accumulation in most plant species, such as vascular plants, green algae, mosses and ferns [11]. The L-galactose biosynthesis pathway starts from D-glucose–6-P and involves a total of nine steps of enzymatic reaction. All the enzymes and the corresponding coding genes involved in this biosynthetic pathway have been identified and well characterized in a lot of higher plants [12] (Fig. 1).

GDP-L-galactose phosphorylase (GGP), which catalyzes the generation of L-galactose–1-P from GDP-L-galactose, is the first committed step in L-galactose biosynthesis pathway of AsA in many plants [12]. The function of GGP was not discovered until 2007, and the gene encoding GGP was the last gene cloned from the L-galactose pathway [13, 14]. Since then, *GGP* genes have been identified and functionally characterized in several plant species, such as kiwifruit [15], apple [16], tomato [17], blueberry [18] and so on. In some plant genomes, GGP proteins are usually encoded by multiple homologous genes, such as two (*VTC2* and *VTC5)* and three (*MdGGP1, MdGGP2* and *MdGGP3)* homologous genes encoding GGP were identified in *Arabidopsis* and apple, respectively [16, 19]. Sequence comparison reveals that VTC2 and VTC5 belong to the histidine triad (HIT) protein superfamily and can specifically catalyze the conversion of GDP-L-galactose to L-galactose–1-phosphate [19]. The expressions of *VTC2* and *VTC5* are regulated by light and could be detected throughout the whole growth and development stages and almost all tissues (root, stem, leaf, flower and silique) of *Arabidopsis thaliana,* and the expression level in green tissues is significantly higher than that in roots [19, 20]. *Arabidopsis* VTC2 and VTC5 are both hydrophilic proteins without transmembrane domains and organelle localization sequence [19]. Sub-cellular localization studies showed that *Arabidopsis* VTC2 and tomato SlGGP exist in cytoplasm and nucleolus, suggesting that plant GGP may be a dual-function protein with enzymatic and regulatory functions [20, 21].

GGP is a critical step in regulating the biosynthesis of AsA in plants, and can control AsA biosynthesis at the transcriptional and translational levels. The expression level of *GPP* gene is closely related to the content of AsA in plants. For example, the content of AsA in kiwifruit leaves and fruits at different development stages was well correlated with the expression of *GGP* gene [15]. Moreover, the transcription level of *SlGGP* was correlated to AsA levels in all tissues of tomato [21], and the higher expression level of *GGP* was also associated with higher AsA content in blueberry [18]. Over-expression of *GGP* can significantly increase the concentrations of AsA in organs such as leaves, fruits and tubers of various

plants, while suppression of *SlGGP* could lead to the decrease of AsA level obviously [15, 21, 22]. These studies suggest that *GGP* is a major control point of AsA biosynthesis in plants. At the translational level, a highly conserved upstream open reading frame (uORF) in the 5' untranslated region (UTR) of *GGP* regulates AsA biosynthesis by forming a feedback loop. The uORF structure regulates the concentration of AsA and the translation of GGP. Under high concentration of AsA, the uORF is translated and inhibits the translation of *GGP,* while under low concentration of AsA, the uORF will not be translated and *GGP* can be smoothly translated to synthesize AsA [23]. Genome editing of uORF of *LsGGP2* can increase AsA concentration significantly in lettuce leaves and thus also increase plant tolerance to oxidative stress [24]. Similar results were also obtained by editing the uORF of *SlGGP1* in tomato [25]. The feedback regulation of AsA biosynthesis suggests that regulation mechanism at translation level also plays an important role in the biosynthesis of AsA.

In view of the important functions of AsA in maintaining normal life activities in almost all living organisms, the AsA biosynthesis pathways and the corresponding structural genes, especially the control points such as the *GME* and *GGP,* have received much attention in recent years. As the first committed step of AsA biosynthesis pathway, *GGP* has attracted particular attention and has been widely investigated. At present, the physical and chemical properties, expression characteristics, roles in plant AsA accumulation and the molecular mechanisms of regulating AsA biosynthesis in plants of this gene have been well understood. However, it is still not known the evolutionary patterns and functional divergence of plant *GGP.* In this study, 149 homologous sequences of *GGP* genes were sampled from 71 plant species representing the major groups of Viridiplantae, and their phylogenetic relationships, gene duplication and molecular evolution analyses were first investigated systematically. The results of this study shed light on the evolutionary patterns of plant *GGP* genes and paved the way to further understand the biological functions of the gene in plant AsA biosynthesis.

Results

Identification of the *GGP* genes in plant kingdom

In order to explore and better understand the evolutionary patterns of plant *GGP* genes, comprehensive homology based BLAST searches were performed. Beside the species available in Phytozome V12.1, the well-known vitamin C rich kiwifruit species, such as *Actinidia chinensis, A.deliciosa, A. eriantha* and *A.rufa,* were also mined from NCBI to identify the homologs of plant *GGP* genes. Furthermore, homologs were also identified from gymnosperm species, including *Gnetum montanum, Picea sitchensis, Picea abies* and *Pinus taeda,* to enrich the representativeness of our sequence data. In total, 149 homologous sequences encoding putative GGPs were mined from 71 Viridiplantae species in the final data. These species, including 15 monocot and 41 dicot angiosperms, 4 gymnosperms, 1 lycophytes, 3 bryophytes and 7 chlorophytes, represented the main lineages of Viridiplantae. The BLAST results also indicated that *GGP* gene exists widely in various plants. The detail information about the plant species and the total *GGP* sequences used in this study were available in Additional file 1 and Additional file 2: Table S1.

A considerably variable number of the *GGP* genes was observed among the tested Viridiplantae specifies (Additional file 2: Table S1). Most plant species in lineages of eudictos, monoctos, gymnopsperms, lycophytes and bryophytes contained at least two homologuesof *GGP,* and the highest copy number of 5 were found in the eudicot species of *Eucalyptus grandis* and the gymnosperm species of *Pinus taeda.* In a few species, especially in the lineage of chlorophytes, only one copy number of *GGP* gene was found. The protein sequence length of plant GGP ranged from 319 to 618 amino acids, and the average length is 438 amino acids (Additional file 2: Table S1).

Recombination test and phylogenetic analysis of plant *GGP* genes

Before carrying out phylogenetic reconstruction, the potential recombination events in the alignment of plant *GGP* coding sequences were firstly screened by GARD method. The result showed that no evidence of recombination was found. Therefore, the alignment of plant *GGP* genes could be directly used to reconstruct phylogenetic relationships and perform molecular evolutionary analysis.

A phylogenetic tree of plant *GGP* was constructed from the alignment of nucleotide sequences using Bayesian method. The resulting Bayesian phylogenetic tree showed that *GGP* genes from angiosperms (including 87 eudicot sequences and 34 monocot sequences) formed a single lineage with high posterior probability support (Fig. 2). Except the bryophyte gene sequences, which were divided into two separate clades, other sequences from gymnosperms, chlorophytes and lycophytes all formed a single lineage with high posterior probabilities, respectively (Fig. 2).

In the angiosperms lineage, one large-scale duplication event was found prior to the radiation of angiosperms, resulting in two subclades of angiosperm 1 (A1) and angiosperm 2 (A2) with posterior probability values more than 0.85, and each of the two subclades contained monocotyledon and dicotyledon *GGP* gene sequences (Fig. 2). Furthermore, another three major duplication events could also be identified in the eudicots 1 of A1 subclade, which occurred before the radiation of brassicaceae, fabaceae and crassulaceae with strongly posterior probability support respectively, leading to the expansion of these three families (Fig. 2). Besides, a major duplication event could also be found within the lineage of gymnosperms with high posterior probability support (Fig. 2). Except the large-scale duplication events, several taxa-specific duplications could also be found in the phylogenetic tree, such as *Ricinus communis, Manihot esculenta, Fragaria vesca, Gossypium raimondii, Eucalyptus grandis, Glycine max, Daucus carota, Mimulus guttatus* in eudicots, *Panicum virgatum, Musa acuminate* and *Spirodela polyrhiza* in monocots, and *Pinus taeda* in gymnosperms. All of these taxa-specific duplications had high posterior probability values more than 0.9 (Fig. 2).

Gene structures and conserved motifs of *GGP* genes

To gain more insight into the structural diversification, conservation and the evolution of plant *GGP* genes, their exon-intron diagrams were generated using GSDS website and illustrated according to their evolutionary relationships (Fig. 3). As the genomic sequences of some genes, included *FvGGP–3* (*Fragaria vesca*)„ *AlGGP–2* (*Arabidopsis lyrate*)„ *AdGGP* (*A. deliciosa*)„ *ArGGP* (*A. rufa*)„ *AeGGP* (*A. eriantha*)„ *PsGGP* (*Picea sithensis)* and *PtGGP–1* (*Picea taeda*)„ were not available at the moment, and thus the exon-intron structure of them were not displayed in this study. As showed in Fig. 3B and 3C, the number of exons varied greatly among different genes, generally ranging from 1 to 11. However, most of the plant *GGP* genes share a similar exon-intron organization, and more importantly, genes within the same lineage usually have the same exon-intron organization. For example, genes within the eudicots 1 lineage varied from 5 to 11 exons, while most of them (73.6%) contained 7 exons. The exon numbers of genes within the lineages of monocots 1, monocots 2, eudicots 2, gymnosperms, chlorophytes, bryophytes and lycophytes contained 6–8, 5–6, 5–6, 7–9, 1–9, 6–8 and 7 exons, respectively. In the lineage monocots 2, all of the genes contained 6 exons except for *SppGGP–1* (*Spirodela polyrhiza*)„ which contained only 5 exons. This may be due to the loss of the fifth intron in *SppGGP–1* gene. Compared with the genes in other lineages, the exons-intron structure of the majority genes in chlorophytes varied greatly, and two intron-less genes and one gene with 9 exons were found in this lineage (Fig. 3B). Furthermore, a large divergence of intron length was observed in a few genes, such as *EgGGP–5* (*Eucalyptus grandis)* and *StGGP–2* (*Solanum tuberosum)* in eudicots 2, *PaGGP–2* (*Picea abies)* in gymnosperms and *DsGGP* (*Dunaliella salina)* in chlorophytes contained several extremely long introns, which were significantly longer than other genes (Fig. 3C).

To investigate the structural divergences and the structural evolution of plant GGP proteins, the conserved motifs was estimated using the MEME online tool. As exhibited in Fig. 3D, a total of 10 conserved motifs were identified and the motifs were present in almost all sequences. Motif compositions and distributions were found to be conserved in most plant GGP proteins sequences, especially within the same lineage members. Some motifs were found to be lacked in a few GGP sequences. For example, motif 1 and 2 were lacked in RcGGP–2 in eudicots 1, SvGGP–2 and SiGGP–2 in monocots 1, and BdGGP–3 in monocots 2, motif 2 and 3 were lacked in BoGGP–2 in eudicots 1, motif 3 and 10 lacked in FgGGP–5 in eudicots 2. It is interesting to note that motif structural and distribution divergences mainly occurred in the lineage of monocots 1 and chlorophytes, especially in chlorophytes where almost all members in this lineage lacked at least one motif (Fig. 3D).

Molecular evolutionary analysis of plant *GGP* genes

Different likelihood-based methods implemented in Codeml program were used to assess the type and strength of natural selection acting on plant *GGP* genes. The branch models were firstly used to test the variation of selective pressure among different branches of the phylogeny tree. The one ratio model M0, which assumes a single ω across all branches and sites in the phylogeny, estimated the ω_0 value for plant *GGP* genes was 0.09302 (Table 1), suggesting that the evolution of *GGP* genes was predominated by strong purifying selection. The free ratio model Mf, which assumes each branch in the tree has an

independent ω value, was compared with M0 using LRT method. And the result showed that Mf model was statistically significant better than M0 model ($-2\Delta lnL = 1072.920122$, $p < 0.0001$) (Table 1), illustrating that the evolutionary rates were different among branches of the phylogenetic tree. A large-scale duplication event was identified in the angiosperm lineage, which gave rise to the angiosperm lineage to split into two sub-lineages of angiosperm 1 (A1) and angiosperm 2 (A2) (Fig. 2). The lineage-specific two-ratio model was employed to detect the changes of selection pressures between different lineages after the duplication event, and the ancestral branches leading to angiosperm, angiosperm 1, angiosperm 2, eudicots 1, monocots 1, eudicots 2 and monocots 2 were set as foreground branch, separately. The results of two-ratio model analyses were given in Table 1. For the ancestral branch leading to angiosperm as foreground branch, the estimated ω value ($\omega_{angiosperm} = 0.06834$) was lower than that of background value ($\omega_0 = 0.09601$), suggesting that the selection pressure has changed during the evolution of angiosperm lineage. However, the LRT statistic result showed that the two-ratio model did not better fit than the null model M0 ($-2\Delta lnL = 2.586622$, $p = 0.1078$) (Table 1), indicating the selection pressure after the duplication event has not changed significantly. For the two sub-lineages after the duplication, the estimated ω value for angiosperm 1 ($\omega_{angiosperm\ 1} = 949.49270$) was larger than that of angiosperm 2 ($\omega_{angiosperm\ 2} = 0.06067$). Both angiosperm 1 and angiosperm 2 lineages were composed of monocotyledon sub-lineage and dicotyledon sub-lineage, respectively. The ω ratios under two-ratio model for eudicot 1 ($\omega_{eudicots\ 1} = 1.65002$) and monocot 1 ($\omega_{monocots\ 1} = 0.20233$) in angiosperm 1 were also larger than eudicot 2 ($\omega_{eudicots\ 2} = 0.08908$) and monocot 2 ($\omega_{monocots\ 2} = 0.02125$) in angiosperm 2, respectively (Table 1), which consisted with the results of angiosperm 1 and angiosperm 2 as mentioned before. For the comparison between the two-ratio model and the one ratio model, only the ancestral branches leading to eudicots 1 and monocots 2 were found significantly different from their background branches. The three ratio model assumed three categories of ω values, corresponding to all branches predating the angiosperm duplication ($\omega 0 = 0.09293$), angiosperm 1 ($\omega_{angiosperm\ 1} = 999.00000$) and angiosperm 2 ($\omega_{angiosperm\ 2} = 0.05179$), respectively. Although the angiosperm 1 had a higher ω than angiosperm 2, the three ratio model did not better fit significantly better than the two ratio model according to the LRT statistic. In general, these results indicated that selection pressures experienced by different lineages were different after the duplication of angiosperm, and *GGP* genes in angiosperm 2 was subjected to more relaxed selection constraints during evolution.

Site-specific codon models were then applied to explore ω value variation across different codon sites and identify potential sites under positive selection. The comparison between M0 and M3 showed that M3 fit the data significantly better than the M0 model ($-2\Delta lnL = 2937.632$, $p < 0.0001$), suggesting that ω values were not homogeneous across different sites. However, the positive selection models of M2a and M8 did not fit the data significantly better than their corresponding negative models of M1a and M7, respectively, and failed to identify any sites under positive selection (Additional file 3: Table S2).

The more powerful branch-site models were also applied to test for episodic positive selection acting on a subset of sites along specific branches. First of all, the main lineages of angiosperm, angiosperm 1, angiosperm 2, eudicots 1, eudicots 2, monocots 1, monocots 2, gymnosperms, chlorophytes, bryophytes

and lycophytes were assigned as foreground branches, respectively. The LRT tests showed that no significant evidence of positive selection were detected in those lineages (Additional file 4: Table S3). Then, to test whether a particular branch in the Bayesian phylogenetic tree was under positive selection, each branch in the phylogenetic tree was assigned as foreground branch and the remaining branches as background branch. The LRT tests detected evidence of positive selection on 22 branches as showed in Additional file 5: Table S4, including *Ananas comosus (AncGGP–3)*„ *Arabidopsis thaliana (AtGGP–1)*„ *Brachypodium stacei* (*GsGGP–1)*„ *Brassica oleracea* (*BoGGP–2)*„ *Brassica rapa (GrGGP–4)*„ *Chlamydomonas reinhardtii* (*ChrGGP)*„ *Coccomyxa subellipsoidea* (*CosGGP)*„ *Dunaliella salina* (*DsGGP)*„ *Eucalyptus grandis* (*EgGGP–5)*„ *Glycine max* (*GmGGP–4)*„ *Marchantia polymorpha* (*MpGGP)*„ *Micromonas pusilla* (*GpGGP)*„ *Micromonas sp RCC299* (*MsGGP)*„ *Mimulus guttatus* (*MgGGP–2)*„ *Musa acuminate (MaGGP–1)*„ *Ostreococcus lucimarinus* (*OlGGP)*„ *Panicum virgatum* (*PvGGP–2)*„ *Picea abies (PaGGP–3)*„ *Selaginella moellendorffii* (*SmGGP–2)*„ *Selaginella moellendorffii* (*SmGGP–3)*„ *Volvox carteri* (*VcGGP)*„ *Zea mays (ZmGGP–1)*.. These positively selected branches were labeled in the phylogenetic tree as showed in Fig. 2, and it can be observed that the angiosperm lineage contained 11 branches under positive selection (five branches in eudicots 1, one branch in monocots 1, one branch in eudicots 2, four in monocots 2), the gymnosperms lineage and the bryophytes lineage each contained only one branch, the lycophytes lineage contained two branches and the chlorophytes lineage contained seven branches. However, only 12 branches, mainly distributed in lineages of eudicots1(two species), monocots 2 (two species), gymnosperms (one species) and chlorophytes (seven species), were supposed to be under positive selection after Bonferroni correction was applied for multiple tests (Fig. 2 and Additional file 5: Table S4). Notably, varying numbers of putative positively selected amino acid sites with posterior probability more than 0.95 under BEB level on these branches were identified as showed in Additional file 5: Table S4.

Discussion

As an enzyme co-factor and antioxidant, AsA plays an important role in plant growth and development, helping plants to against abiotic and biotic stress and adapt to diverse environmental conditions. AsA is widely found in almost all plants, while the concentration of AsA varies considerably across different plant species. For example, AsA concentrations in higher plants range approximately from 2 to 135 μmol g^{-1} FW (fresh weight), however, green algae species of *Ulva compressa* and bryophyte species of *Hypnum plumaeforme* exhibit AsA concentrations of about 0.5 μmol g^{-1} FW and 0.1—0.6 μmol g^{-1} FW, respectively [26, 27]. Plant AsA level is closely related to external environmental factors such as light intensity, UV-B and temperature, and is controlled by internal metabolism mechanisms including biosynthesis, recycling and degradation. L-galactose pathway, which is the principal route of AsA biosynthesis, contributes a lot in maintaining AsA pools in most plants. As a rate-limiting step in L-galactose pathway in both green algae and higher plants, *GGP* plays an essential role in plant AsA biosynthesis and the expression level of *GGP* largely determines the synthesis rate of AsA [28]. In this study, 147 sequences of *GGP* homologs were retrieved from 71 plant species, which representing major

Viridiplantae lineages including eudicots, monocots, gymnosperms, lycophytes, bryophytes and chlorophytes, and the functional diversity and evolutionary patterns were systematically explored.

Plant *GGP* gene has undergone several duplication events during evolution. 50 out of 71 species collected in this study, which mainly located in lineages of angiosperms and gymnosperms, contained more than two copies of *GGP,* and the species containing only one copy of *GGP* gene were mainly located in the lineage of chlorophytes. Phylogenetic analyses revealed five well-supported major duplication events in the evolutionary history of plant *GGP* genes. Gene duplication usually comes from whole genome duplication (WGD) events, which leads to an increase in the number of gene copies. WGD events have been commonly detected in many lineages of eukaryotes, especially in angiosperms. WGDs occurred many times during the evolution of angiosperms, which greatly promoted the adaptive radiation of angiosperms [29]. Among the five major duplication events identified in the study, four duplication events occurred in the lineage of angiosperms and coincided with WGD events previously identified in angiosperms. The first major duplication even occurred in the angiosperm ancestral species, resulting in two sub-lineages of angiosperm 1 and angiosperm 2 (Fig. 2). The other three duplication events were Brassicaceae, Fabaceae and Crassulaceae specific, respectively, and all occurred in eudicots 1 in the sub-lineage of angiosperm 1 (Fig. 2). The three gene duplication events coincided with WGD events in the Brassicales [30, 31], Fabaceae [32–34] and Crassulaceae [35], respectively. The last major gene duplication event was identified in the lineage of gymnosperms, but occurred limited to Pinaceae. This result was consistent with previous studies on early genome duplications in gymnosperms, that is, WGD events were detected in Pinaceae and other gymnosperms, while no evidence of WGDs were detected in the genome of gnetophytes [36, 37]. Moreover, a number of species-specific duplications were also identified frequently in the seed plant lineages (Fig. 2). In general, AsA content has a tendency to increase during the evolution of plants, that is, the concentration of AsA in higher plants is usually much higher than that in bryophytes and green algae [26, 28]. The five major duplication events and a number of taxon-specific duplications led to the rapid expansion of *GGP* genes in seed plants, especially in angiosperms. This may be one of the reasons for the higher AsA content in angiosperms.

Most plant *GGP* genes have similar exon-intron structure and relatively conservative motif composition and distribution. The structure of *GGP* gene was more conserved in land plants, but varied greatly in green algae, indicating that *GGP* may have undergone great differentiation in the early stages of plant evolution. Most GGP proteins had a conserved motif arrangement and composition, suggesting that plant GGPs have similar catalytic functions. Nevertheless, there may be some differences in the functions of GGP homologues in the same plant. For example, *VTC2* and *VTC5* are the two homologous genes encoding GGP in *Arabidospsis thaliana,* while their expression level and expression tissues were somewhat different. *VTC2* seemed to plays a more important role in AsA biosynthesis [19]. Studies in tomatoes showed that although *SlGGP2* played a role in regulating the concentration of AsA in fruit, the expression level of *SlGGP1* was more closely related to the level of AsA during fruit ripening [38]. Studies on the *LsGGP1* and *LsGGP2* uORF mutants in lettuce also revealed functional differences between the two isozymes, suggesting that *LsGGP2* may be the major GGP isoenzyme that regulates AsA

biosynthesis [24]. The functional difference between plant *GGP* homologous genes may be related to gene duplication, which usually leads to subfunctionalization or neofunctionalization [39].

GGP is generally considered as a major determinant gene in plant AsA biosynthesis, and plays an important role in regulating AsA concentrations in many plants. In this study, evolutionary analysis revealed that plant *GGP* gene was mainly restricted by strong purifying selection (ω_0 = 0.09302), which indicated the functional importance and conservativeness of plant *GGP* genes during evolution. The molecular evolutionary results of *GGP* were similar to that of *GME,* which is the upstream gene of *GGP* in L-galactose pathway and is also considered as a key gene in plant AsA biosynthesis, and also had undergone strong purifying selection during evolution (ω_0 = 0.0287) [27]. Moreover, a total of 22 branches were identified under positive selection. Even after Bonferroni correction, there were still 12 branches under positive selection, most of which (seven branches) were in the chlorophytes lineage. These results were also consistent with the results in the *GME,* where most of the positively selected branches detected in the *GME* species were located in the green algae lineage [27], and also suggesting that the evolutionary innovation of *GGP* genes also mainly occurred in the early stages of plant evolution and played an important role in helping plants adapt to new and challenging environments.

In plants, the L-galactose pathway involves nine consecutive enzymes, of which *GME* and *GGP* are considered to be the critical steps to regulate the synthesis of AsA. The expression of *GME* and *GGP* is induced by light and abiotic stress, and these two genes operate synergistically to regulate AsA biosynthesis [12, 40]. Although both *GME* and *GGP* genes were under strong purifying selection, the upstream gene *GME* (ω_0 = 0.0287) was subject to more selective constraints than the downstream gene *GGP* (ω_0 = 0.09302). The results seemed to consistent with the hypothesis that upstream genes in metabolic pathways are more constrained than downstream genes [41, 42]. At present, only the evolutionary patterns of *GME* and *GGP* have been studied, while the selection signatures of other genes in L-galactose pathway are not still clear, and the factors affecting the evolution rate of genes in L-galactose pathway are also uncertain. Molecular evolution studies of other genes in the L-galactose pathway in future works will help to clarify the evolution patterns of the L-galactose pathway genes and identify factors affecting the selection pressure differences among the pathway genes.

Conclusion

In conclusion, the molecular evolutionary pattern of plant *GGP* genes, which play a key regulatory role in AsA biosynthesis, were first systematically explored in this study. Most plant *GGP* genes had similar gene structure and motif patterns, indicating that plant *GGP* genes have conservative functions. Molecular evolutionary studies showed that *GGP* genes were mainly constrained by strong purifying selection, which indicated the functional importance of *GGP.* A few branches were identified under positive selection and most of which located in the chlorophytes lineage, indicating that episodic diversifying selection played a role during the evolution of plant *GGP* genes. Several major duplication events and taxon-specific duplication events were identified in seed plants, especially in angiosperm lineages, which promoted the radiation of *GGP* gene in angiosperms. The conservative function of *GGP* gene and its

rapid expansion in angiosperms may be one of the reasons for the increase of AsA content in angiosperms, enabling angiosperms to adapt to changing environments.

Methods

Acquisition and characterization of plant *GGP* coding sequences

To retrieve potential plant *GGP* coding sequences, multiple genomic databases were searched using BLAST method. The amino acid sequences, genomic sequences and coding DNA sequences (CDS) of plant *GGP* involved in this study were mainly collected from online databases of Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html) and National Center of Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). In order to identify *GGP* coding sequences in Viridiplantae, the *Arabidopsis thaliana* GGPamino acid sequences of VTC2 (At4g26850) and VTC5 (At5g55120), which were downloaded from the TAIR database (https://www.arabidopsis.org/), were used as queries to carry out BLASTP searches against the databases of Phytozome v12.1 and NCBI with default algorithm parameters. In addition, to obtain the *GGP* coding sequences from gymnosperm lineage, BLASTP searches were also performed against the genomes of *Picea abies* and *Picea taeda* in ConGenIE database (http://congenie.org/citation) using protein sequences of VTC2 and VTC5. Besides, the genome sequence of *Gnetum montanum* was downloaded from DRYAD website (https://datadryad.org/handle/10255/dryad.186891) and searched for *GGP* orthologs by local BLASTP method using the two Arabidopsis GGP proteins as queries. All identical, redundant, partial and incomplete sequences were manually identified and eliminated from the original sequences, and only the full-length coding sequences were retained in the final dataset.

Multiple sequence alignment

Amino acid sequences of the collected plant GGP were firstly aligned using MAFFT program v7.158 [43] with default parameters. After manually curated in BioEdit v7.1.13 software [44], the multiple sequence alignment of the amino acid sequences was uploaded to PAL2NAL website (http://www.bork.embl.de/pal2nal/) [45] and then converted into the corresponding coding sequence alignment. Subsequently, the codon alignment was filtered using the program Gblocks v0.91b [46] to trim ambiguously aligned positions and to obtain conserved regions, with 50% gapped positions allowed and all other parameters were kept at default options.

Detection of recombination events

It is well known that recombination events may adversely affect the accuracy and efficiency of phylogenetic reconstruction and molecular evolutionary analysis [47–49]. As a result, to avoid the

potential impact of recombination on our dataset of plant *GGP* protein-coding DNA sequences, the GARD recombination detection method [50] implemented in Datamonkey web-server (http://www.datamonkey.org/) [51] was initially utilized to screen for evidence of recombination breakpoints prior to phylogenetic and evolutionary analyses.

Phylogenetic tree reconstruction

The nucleotide phylogenetic tree of plant GGPs were generated by Bayesian inference implemented in the program MrBayes v3.2.6 [52]. Prior to reconstruct the Bayesian phylogeny, the best-fit nucleotide substitution model of GTR+I+G was determined using MrModeltest v2.3 under the standard of Akaike Information Criterion (AIC) [53]. The Bayesian phylogenetic reconstruction was run for 10,000,000 Markov Chain Monte Carlo (MCMC) generations and sampled every 100 generations. Trees from the first twenty-five percent of the sampled generations were discarded as burn-in. The final phylogenetic tree was manipulated and visualized using iTOL web server (https://itol.embl.de/) [54].

Analysis of exon-intron structure and conserved motifs

The Gene Structure Display Server v2.0 (GSDS) (http://gsds.cbi.pku.edu.cn/) [55] online tool was employed to display the exon-intron structure features of plant *GGP* genes by comparing the full-length CDS sequences with their corresponding genomic sequences. Moreover, the motif analysis tool of Multiple Em for Motif Elicitation v5.0.5 (MEME) (http://meme-suite.org/tools/meme) [56] was used to detect conserved motif structures of plant GGP protein sequences with mostly default parameters except for the number of motifs was set to 10.

Molecular evolutionary analyses

To test for signatures of positive selection in plant *GGP* genes, several codon-based maximum-likelihood models implemented in CODEML program in the PAML package v4.9i [57] were used in this study. And the aligned codon-based sequences and the reconstructed phylogenetic tree were fed into the CODEML program to estimate the nonsynonymous (d_N) versus synonymous substitution (d_S) rate ratio ($\omega = d_N/d_S$). The ω values estimated by the maximum likelihood methods is a useful measurement to identify adaptive molecular evolution, with $\omega = 1$, < 1, and > 1 meaning neutral evolution, negative selection, and positive selection, respectively [58].

To test the variation of ω between amino acid sites and identify potential sites evolving by positive selection, three pairs of site-specific models were compared, including M0 (one-ratio model) versus M3 (discrete model), M1a (nearly neutral model) versus M2a (positive selection model), and M7 (neutral, beta model) versus M8 (selection, beta & ω model) [58]. The M0 assumes a constant ω ratio for all sites and all branches, whereas M3 assumes all the sites in a branch have a different ω ration. The neutral models

M1a and M7 restricts sites with ω < 1, while selection models M2a and M8 add an additional class of sites with ω > 1. The comparison of the three pairs of models was performed through Likelihood Ratio Test (LRT) with chi-square (χ^2) distribution. If the LRT was significant (*p*-value < 0.01), then the Bayes Empirical Bayes (BEB) [59] approach was employed to identify amino acid sites under positive selection (posterior probability ≥ 90%)

To examine the variation of ω among different branches of the phylogenetic tree, the branch models of M0 (one-ratio model) and Mf (free ratio model) were compared through LRT method. The free-ratio model Mf assumes each branch of the tree has an independent ω value, and the corresponding one-ratio model, which assumes a constant ω value among all branches, was used as the null expectation [60]. Then, the two ratio model and the three ratio model were employed to evaluate selection pressures acting upon interesting lineages, especially in lineages experienced duplication events, such as the lineage of angiosperm. Besides, the improved branch-site models in CODEML were also used to detected signals of positive selection along particular branches [61]. For branch model (two ratio model and three ratio model) and branch-site model analyses, the lineages or branches of interest were prespecified as foreground branches that allow positive selection, while the rest of lineages or branches were defined as background branches that allow negative or neutral selection. The LRT test was also used to identify signals of natural selection in the foreground branches when performing these models. In addition, the Bonferroni's correction was employed to control the family-wise error rate when multiple branches on the phylogeny were used to detect positive selection in the branch-site test [62].

Abbreviations

AIC: Akaike Information Criterion; AsA: Ascorbic acid; BEB: Bayes Empirical Bayes; CDS: Coding DNA sequences; GGP: GDP-L-galactose phosphorylase; GME: GDP-D-mannose epimerase; GuLO: L-gulonolactone oxidase; GSDS: Gene Structure Display Server; HIT: Histidine triad; LRT: Likelihood Ratio Test; MCMC: Markov Chain Monte Carlo; MEME: Multiple Em for Motif Elicitation; NCBI: National Center of Biotechnology Information; ROS: Reactive oxygen species; UTR: Untranslated region; uORF: Upstream open reading fame; Vc: Vitamine C; WGD: Whole genome duplication;

Declarations

Acknowledgements

The authors are grateful for the grants from National Natural Science Foundation of China and the supports from Department of Horticulture, College of Agronomy, Jiangxi Agricultural University.

Authors' contributions

JT and CH conceived and designed the study. JT and ZH collected the data. JT, ZH and CH performed the data analyses. JT wrote the manuscript. All authors read and approved the final version of the manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (31760567, 31460505), the science and technology research project of the education department of Jiangxi Province (Gjj150382) and the key projects of Weinan Normal University (16ZRRC01).

Availability of data and materials

The amino acid sequences, genomic sequences and coding DNA sequences (CDS) of plant *GGP* were mainly downloaded from online databases, including Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html), NCBI (https://www.ncbi.nlm.nih.gov/), ConGenIE database (http://congenie.org/citation) and DRYAD website (https://datadryad.org/handle/10255/dryad.186891). The detail information of plant *GGP* genes involved in this study is shown in Additional file 2: Table 1.

Ethics approval and consent to participate

No animals were used in this study.

Consent of publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] College of Agronomy, Jiangxi Agricultural University, Nanchang, China

[2] Institute of Kiwifruit, Jiangxi Agricultural University, Nanchang, China

[3] College of Chemistry and Materials, Weinan Normal University, Weinan, China

References

1.Macknight RC, Laing WA, Bulley SM, Broad RC, Johnson AA, Hellens RP. Increasing ascorbate levels in crops to enhance human nutrition and plant abiotic stress tolerance. Curr Opin Biotechnol.2017;44:153–60.

2.Mellidou I, Koukounaras A, Chatzopoulou F, Kostas S, Kanellis AK. Plant vitamin C: One single molecule with a plethora of roles. In: Yahia EM, editor. Fruit and vegetable phytochemicals: Chemistry and human health.Volume 1, 2nd ed. John Wiley & Sons, Ltd. 2017. p. 463–98.

3.Gallie DR. L-ascorbic acid: a multifunctional molecule supporting plant growth and development. Scientifica.2013;2013:795964.

4.Fenech M, Amaya I, Valpuesta V, Botella MA. Vitamin C content in fruits: Biosynthesis and regulation. Front Plant Sci.2018;9:2006.

5.Camarena V, Wang G. The epigenetic role of vitamin C in health and disease. Cell Mol Life Sci.2016;*73:*1645–58.

6.Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi KC. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. J Biol Chem.1994;269:13685–88.

7.Wheeler GL, Jones MA, Smirnoff N. The biosynthetic pathway of vitamin C in higher plants. Nature.1998;393:365–69.

8.Wolucka BA, Van Montagu M. GDP-mannose 3', 5'-epimerase forms GDP-L-gulose, a putative intermediate for the de novo biosynthesis of vitamin C in plants. J Bio Chem.2003; 278:47483–90.

9.Agius F, González-Lamothe R, Caballero JL, Muñoz-Blanco J, Botella MA, Valpuesta V. Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. Nat Biotechnol. 2003;21:177–81.

10.Lorence A, Chevone BI, Mendes P, Nessler CL. *myo*-Inositol oxygenase offers a possible entry point into plant ascorbate biosynthesis. Plant Physiol.2004;134:1200–5.

11.Ishikawa T, Maruta T, Yoshimura K, Smirnoff N. Biosynthesis and regulation of ascorbic acid in plants. In: Gupta D, Palma J, Corpas F, editors.Antioxidants and Antioxidant Enzymes in Higher Plants. Cham: Springer; 2018. p. 163–79.

12.Bulley S, Laing W. The regulation of ascorbate biosynthesis. Curr Opin Plant Biol.2016;33:15–22.

13.Linster CL, Gomez TA, Christensen KC, Adler LN, Young BD, Brenner C, Clarke SG. *Arabidopsis VTC2* encodes a GDP-L-galactose phosphorylase, the last unknown enzyme in the Smirnoff-Wheeler pathway to

ascorbic acid in plants. J Bio Chem.2007;282:18879–85.

14.Laing WA, Wright MA, Cooney J, Bulley SM. The missing step of the L-galactose pathway of ascorbate biosynthesis in plants, an L-galactose guanyltransferase, increases leaf ascorbate content. Proc Natl Acad Sci USA.2007;104:9534–39.

15.Bulley SM, Rassam M, Hoser D, Otto W, Schunemann N, Wright M, et al. Gene expression studies in kiwifruit and gene over-expression in *Arabidopsis* indicates that GDP-L-galactose guanyltransferase is a major control point of vitamin C biosynthesis. J Exp Bot. 2009;60:765–78.

16.Mellidou I, Chagne D, Laing WA, Keulemans J, Davey MW. Allelic variation in paralogs of GDP-L-galactose phosphorylase is a major determinant of vitamin C concentrations in apple fruit. Plant Physiol.2012;160:1613–29.

17.Wang L, Meng X, Yang D, Ma N, Wang G, Meng Q. Overexpression of tomato GDP-L-galactose phosphorylase gene in tobacco improves tolerance to chilling stress. Plant Cell Rep.2014;33:1441–51.

18.Liu F, Wang L, Gu L, Zhao W, Su H, Cheng X. Higher transcription levels in ascorbic acid biosynthetic and recycling genes were associated with higher ascorbic acid accumulation in blueberry. Food Chem.2015;188:399–405.

19.Dowdle J, Ishikawa T, Gatzek S, Rolinski S, Smirnoff N. Two genes in *Arabidopsis thaliana* encoding GDP-L-galactose phosphorylase are required for ascorbate biosynthesis and seedling viability. Plant J.2007;52:673–89.

20.Muller-Moule P. An expression analysis of the ascorbate biosynthesis enzyme VTC2. Plant Mol Biol. 2008;68:31–41.

21.Wang LY, Li D, Deng YS, Lv W, Meng QW. Antisense-mediated depletion of tomato GDP-L-galactose phosphorylase increases susceptibility to chilling stress. J Plant Physiol.2013;170:303–14.

22.Bulley S, Wright M, Rommens C, Yan H, Rassam M, Lin-Wang K, et al. Enhancing ascorbate in fruits and tubers through over-expression of the L-galactose pathway gene GDP-L-galactose phosphorylase. Plant Biotechnol J.2012;10:390–7.

23.Laing WA, Martinez-Sanchez M, Wright MA, Bulley SM, Brewster D, Dare AP, et al. An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in *Arabidopsis.*Plant Cell.2015;27:772–86.

24.Zhang H, Si X, Ji X, Fan R, Liu J, Chen K, et al. Genome editing of upstream open reading frames enables translational control in plants. Nat Biotechnol.2018;36:894–8.

25.Li T, Yang X, Yu Y, Si X, Zhai X, Zhang H, et al. Domestication of wild tomato is accelerated by genome editing. Nat Biotechnol.2018;36:1160–3.

26.Gest N, Gautier H, Stevens R. Ascorbate as seen through plant evolution: the rise of a successful molecule? J Exp Bot.2013;64:33–53.

27.Tao J, Wu H, Li Z, Huang C, Xu X. Molecular evolution of GDP-D-mannose epimerase (*GME),,* a key gene in plant ascorbic acid biosynthesis. Front Plant Sci.2018;9:1293.

28.Vidal-Meireles A, Neupert J, Zsigmond L, Rosado-Souza L, Kovacs L, Nagy V, et al. Regulation of ascorbate biosynthesis in green algae has evolved to enable rapid stress-induced response via the *VTC2* gene encoding GDP-L-galactose phosphorylase. New Phytol. 2017; 214:668–81.

29.Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant.2018;11:414–28.

30.Barker MS, Vogel H, Schranz ME. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol Evol. 2009;1:391–9.

31.Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana.* BMC Evol Biol. 2011;11:47.

32.Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83.

33.Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, et al. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. 2011; 480:520–24.

34.Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula.* BMC Genomics. 2014; 15:312.

35.Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, et al. The Kalanchoe genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. Nat Commun. 2017;8:1899.

36.Wan T, Liu ZM, Li LF, Leitch AR, Leitch IJ, Lohaus R, et al. A genome for gnetophytes and early evolution of seed plants. Nat Plants. 2018;4:82–9.

37.Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. Sci Adv. 2015;1:e1501084.

38.Mellidou I, Keulemans J, Kanellis AK, Davey MW. Regulation of fruit ascorbic acid concentrations during ripening in high and low vitamin C tomato cultivars. BMC Plant Biol.2012;12:239.

39.Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73.

40.Mellidou I, Kanellis AK. Genetic control of ascorbic acid biosynthesis and recycling in horticultural crops. Front Chem. 2017;5:50.

41.Clotault J, Peltier D, Soufflet-Freslon V, Briard M, Geoffriau E. Differential selection on carotenoid biosynthesis genes as a function of gene position in the metabolic pathway: a study on the carrot and dicots. PloS ONE. 2012;7:e38724.

42.Rausher MD, Lu Y, Meyer K. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 2008;67:137–44.

43.Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–80.

44.Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999;95–8.

45.Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–612.

46.Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.

47.Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol.2002;54:396–402.

48.Anisimova M, Nielsen R, Yang Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 2003;164:1229–36.

49.Shriner D, Nickle DC, Jensen MA, Mullins JI. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res. 2003;81:115–21.

50.Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. Bioinformatics.2006;22:3096–8.

51.Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol. 2018;35:773–7.

52.Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42.

53.Nylander J. MrModeltest v2.3. Program distributed by the author. Evolutionary Biology Centre, Uppsala University. 2008. website: https://github.com/nylander/MrModeltest2.

54.Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.

55.Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015;31:1296–7.

56.Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res.2009;37:W202–8.

57.Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol.2007;24:1586–91.

58.Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics.2000;155:431–49.

59.Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22:1107–18.

60.Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 199;15:568–73.

61.Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22:2472–9.

62.Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 2007;24:1219–28.

Tables

Table 1. PAML branch model analyses to test the variable selective pressure among branches and after gene duplication

Model	Np[a]	lnL[b]	Parameter estimates	Models compared	d.f.[c]	-2ΔlnL[d]	*p*-value
A: One ratio model M0	297	-49065.194200	$\omega_0 = 0.09302$				
B: Two ratios (angiosperm)	298	-49063.900889	$\omega_0 = 0.09333$, $\omega_{angiosperm} = 0.04308$	B vs. A	1	2.586622	0.1078
C: Two ratios (angiosperm 1)	298	-49063.877071	$\omega_0 = 0.09272$, $\omega_{angiosperm1} = 949.49270$	C vs. A	1	2.634258	0.1046
D: Two ratios (angiosperm 2)	298	-49064.939231	$\omega_0 = 0.09317$, $\omega_{angiosperm2} = 0.06067$	D vs. A	1	0.509938	0.4752
E: Two ratios (eudicots 1)	298	-49061.885830	$\omega_0 = 0.09241$, $\omega_{eudicots1} = 1.65002$	E vs. A	1	6.61674*	0.0101
F: Two ratios (monocots 1)	298	-49063.571050	$\omega_0 = 0.09264$, $\omega_{monocots1} = 0.20233$	F vs. A	1	3.2463	0.0716
G: Two ratios (eudicots 2)	298	-49065.190239	$\omega_0 = 0.09304$, $\omega_{eudicots2} = 0.08908$	G vs. A	1	0.007922	0.9291
H: Two ratios (monocots 2)	298	-49060.105366	$\omega_0 = 0.09359$, $\omega_{monocots2} = 0.02125$	H vs. A	1	10.177668**	0.0014
I: Three ratios (angiosperm 1, angiosperm 2)	299	-49063.354842	$\omega_0 = 0.09293$, $\omega_{angiosperm1} = 999.00000$,	I vs. B	1	1.092094	0.296

			$\omega_{angiosperm2}$ = 0.05179				
J: Free ratio model Mf	591	-48528.734139		J vs. A	294	1072.920122***	< 0.0001

[a] Np: number of esiad parmeters.

[b] lnL: log-likelihood scores.

[c] d.f.: degree of freedom.

[d] -2ΔlnL: twice the log-likelihood difference of the models being compared.

* Significant at $p < 0.05$; ** Significant at $p < 0.01$; ;*** Significant at $p < 0.0001$

Figures

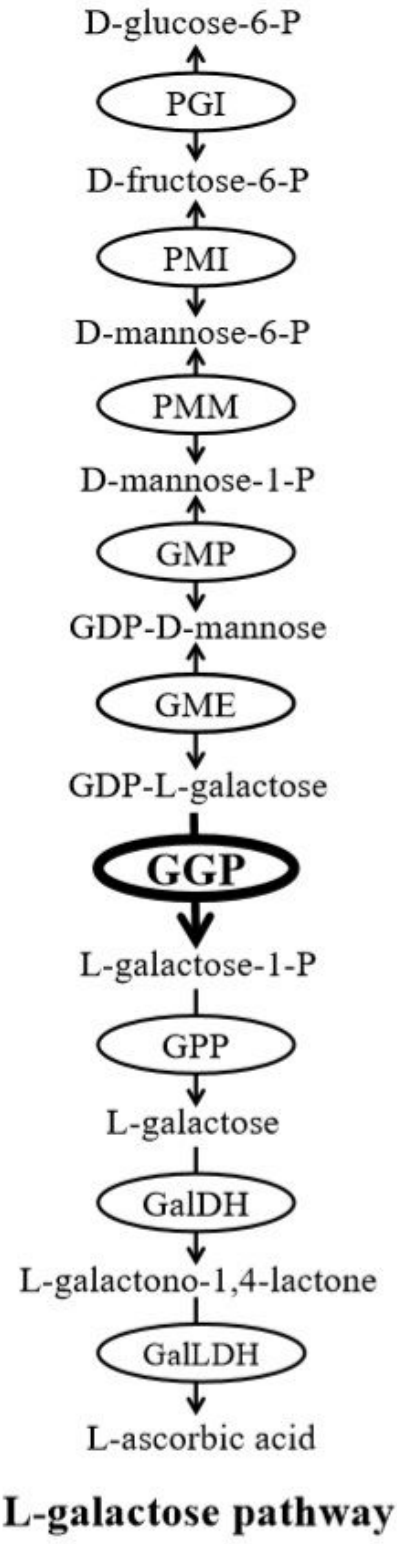

Figure 1

AsA biosynthesis by the L-galactose pathway in plants. Enzymes involved in L-galactose pathway are labeled in the circles, including (1) PGI: glucose-6-phosphate isomerase; (2) PMI: mannose-6-phosphate isomerase; (3) PMM: phosphomannomutase; (4) GMP: GDP-mannose pyrophosphorylase; (5) GME: GDP-mannose-3', 5'-epimerase; (6) GGP: GDP-L-galactose phosphorylase; (7) GPP: L-Galactose-1-phosphate

phosphatase; (8) GalDH: L-Galactose dehydrogenase; (9) GalLDH: L-Galactono-1,4-lactone dehydrogenase.

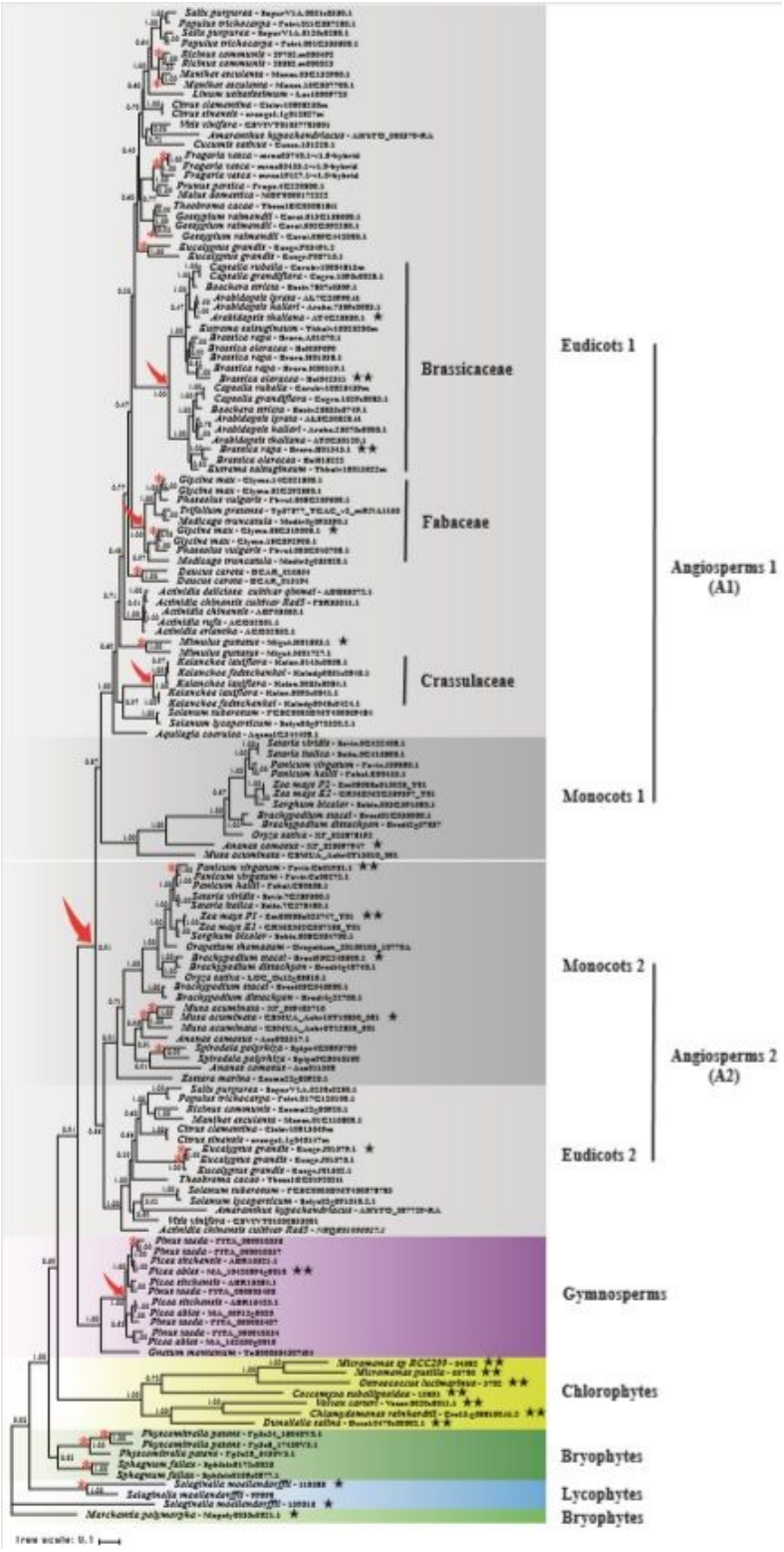

Figure 2

Phylogenetic analyses of plant GGP genes using the Bayesian method. The phylogenetic tree was constructed through the Bayesian method under the GTR+I+G model. Posterior probabilities are labeled near the nodes. The accession number of the GME gene is listed after the name of the species. Red

arrows indicate major duplication events. Red asterisks (□) indicate taxon-specific gene duplications. Black star (□) indicates the branch is identified under episodic diversifying selection by branch-site model. Double black stars (□□) indicate that the branch is still under positive selection after Bonferroni correction.

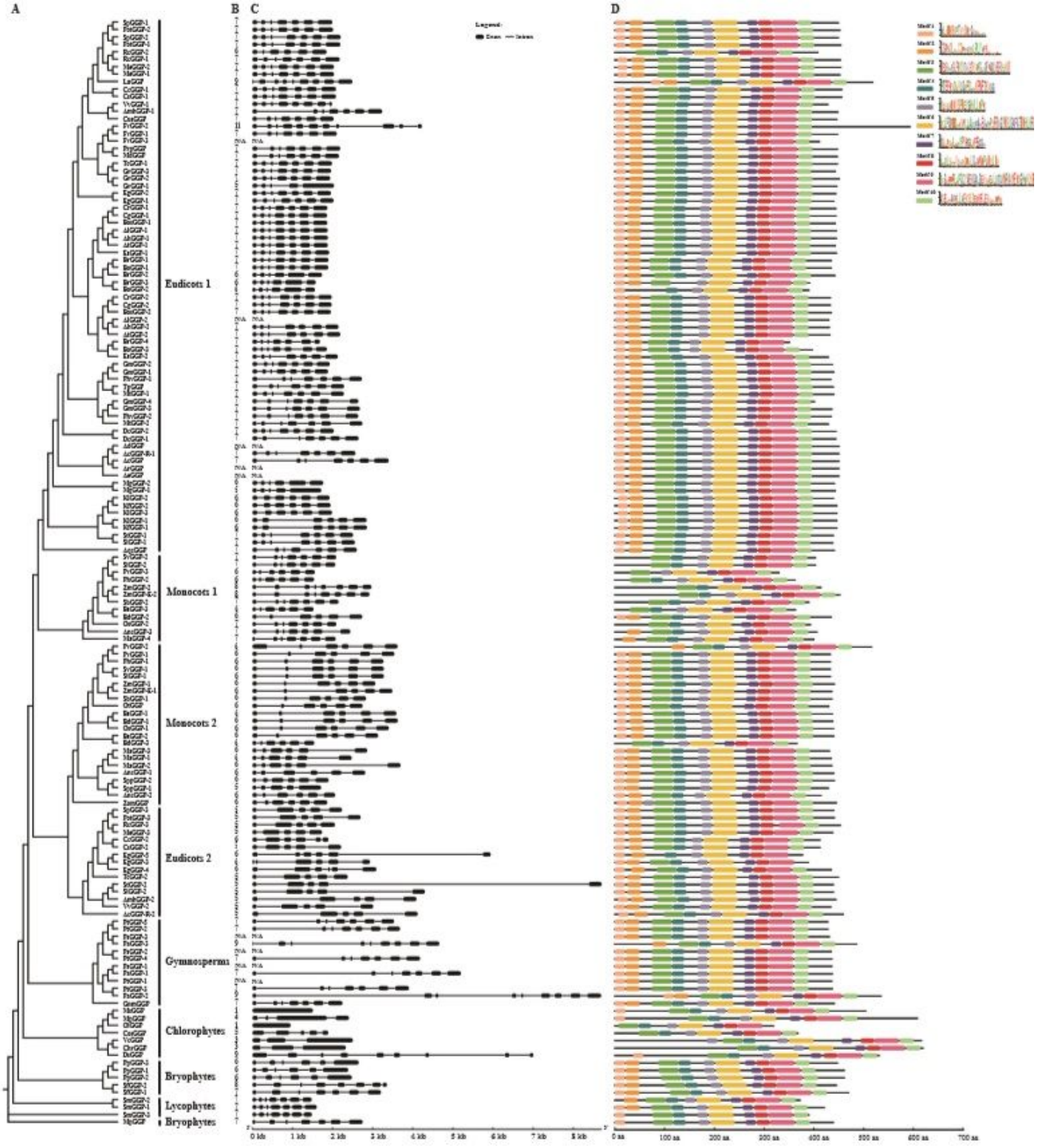

Figure 3

Phylogenetic relationships, gene structures and conserved protein motifs of plant GGP genes. (A) The Bayesian phylogenetic tree of plant GGP genes. (B) Exon number of corresponding GGP genes. (C) Exon-

intron structure of plant GGE genes. (D) The conserved motif composition and distribution of plant GGP proteins. The conserved motifs are displayed in different colored boxes, and the sequence information for each motif is displayed in the form of seqlogo.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile3TableS2.docx
- Additionalfile5TableS4.docx
- Additionalfile1.pdf
- Additionalfile2TableS1.xlsx
- Additionalfile4TableS3.docx