frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 27 May 2014
doi: 10.3389/fpls.2014.00230

Distinct evolutionary strategies in the GGPPS family from plants

Diana Coman*[1], *Adrian Altenhoff*[2,3], *Stefan Zoller*[2,3], *Wilhelm Gruissem*[1] and *Eva Vranová*[1,4]

[1] *Department of Biology, ETH Zurich, Zurich, Switzerland*
[2] *Department of Computer Science, ETH Zurich, Zurich, Switzerland*
[3] *Swiss Institute of Bioinformatics, Zurich, Switzerland*
[4] *Institute of Biology and Ecology, Pavol Jozef Šafárik University, Košice, Slovakia*

Edited by:
Catherine Anne Kidner, University of Edinburgh, UK

Reviewed by:
Jinling Huang, East Carolina University, USA
Ben Holt, University of Oklahoma, USA

****Correspondence:***
Eva Vranová, Faculty of Science, Institute of Biology and Ecology, Pavol Jozef Šafárik University in Košice, Mánesova 23, Košice, 04154, Slovakia
e-mail: eva.vranova@upjs.sk

Multiple geranylgeranyl diphosphate synthases (GGPPS) for biosynthesis of geranylgeranyl diphosphate (GGPP) exist in plants. GGPP is produced in the isoprenoid pathway and is a central precursor for various primary and specialized plant metabolites. Therefore, its biosynthesis is an essential regulatory point in the isoprenoid pathway. We selected 119 GGPPSs from 48 species representing all major plant lineages, based on stringent homology criteria. After the diversification of land plants, the number of *GGPPS* paralogs per species increases. Already in the moss *Physcomitrella patens*, GGPPS appears to be encoded by multiple paralogous genes. In gymnosperms, neofunctionalization of GGPPS may have enabled optimized biosynthesis of primary and specialized metabolites. Notably, lineage-specific expansion of GGPPS occurred in land plants. As a representative species we focused here on *Arabidopsis thaliana*, which retained the highest number of GGPPS paralogs (twelve) among the 48 species we considered in this study. Our results show that the *A. thaliana GGPPS* gene family is an example of evolution involving neo- and subfunctionalization as well as pseudogenization. We propose subfunctionalization as one of the main mechanisms allowing the maintenance of multiple *GGPPS* paralogs in *A. thaliana* genome. Accordingly, the changes in the expression patterns of the *GGPPS* paralogs occurring after gene duplication led to developmental and/or condition specific functional evolution.

Keywords: GGPPS, isoprenoids, paralogs, specialized metabolism, subfunctionalization

INTRODUCTION

Isoprenoids represent the largest group of biologically active specialized metabolites in plants. Many have roles in protecting the plants against pathogens and herbivores or conversely they attract pollinators and seed-dispersing animals. (Bouvier et al., 2005). Other isoprenoids have important roles in photosynthesis and respiration or as hormones (abscisic acid, brassinosteroids, cytokinins, gibberellic acid, strigolactones) in development and growth regulation (Bouvier et al., 2005; Liang, 2009; Vranová et al., 2012).

In spite of their broad diversity of functions and structures, the biosynthesis of all isoprenoids in plants invariably requires two five-carbon (C5) building units: the isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) (Liang et al., 2002; Hsieh et al., 2011; Vranová et al., 2013). In plants, the mevalonic acid pathway (MVA) produces cytosolic IPP, and the methylerythritol pathway (MEP) produces IPP and DMAPP in plastids (Goldstein and Brown, 1990; Rohmer, 1999; Rodríguez-Concepción and Boronat, 2002). The MVA and MEP pathways are linear step enzymatic reactions until the synthesis of the allylic prenyl diphosphates. Then, prenyl diphosphate synthases catalyze chain elongation reactions by coupling IPP to DMAPP producing allylic prenyl diphosphates of different length (Vranová et al., 2013). Most of the essential plant isoprenoids are derived from the C15 and C20 allylic prenyl diphosphates farnesyl-PP (FPP) and geranylgeranyl-PP (GGPP), whose pools represent nodes of the major metabolic branch points in the isoprenoid synthesis (Vranová et al., 2011).

In plants, the enzymes catalyzing the steps upstream of GGPP biosynthesis are encoded either by single copy genes or by pairs of genes (Goldstein and Brown, 1990; Rodríguez-Concepción and Boronat, 2002; Closa et al., 2010; Vranová et al., 2013). Intriguingly, at the GGPP branch point, a high number of genes encoding GGPP synthase is predicted for plant genomes, reaching up to 12 members per species (PLAZA, http://bioinformatics.psb.ugent.be/plaza/).

Multiple gene copies result from duplication events, which can involve individual genes, chromosomal segments, or entire genomes (whole-genome duplication, WGD). Such genes descend from a common ancestor and are homologous (Innan and Kondrashov, 2010). Homologous genes are further classified into paralogs, which are related by duplication events and orthologs, which are genes in different species that evolved from a common ancestor through speciation events (Fitch, 1970). Whereas orthologs tend to share similar functions, paralogs tend to have different roles (Studer and Robinson-Rechavi, 2009). Following duplication, one of the outcomes for a paralog is to accumulate inactivating mutation and become a pseudogene (Innan and Kondrashov, 2010). Alternatively, paralogs are preserved in the genome, particularly if they confer selective advantages. For example, one gene may retain the ancestral function whereas the other undergoes accelerated evolution to acquire a new function ("neofunctionalization") (Innan and Kondrashov, 2010). Or both paralogous copies might specialize and retain only distinct subsets of the ancestral gene function ("subfunctionalization"), which may increase the fitness of the organism (Lynch and Conery, 2000; Lynch and Force, 2000).

Although biosynthesis of GGPP is an essential step in the isoprenoid pathway providing the common precursor for key metabolic pathways involved in both primary and specialized metabolism, to date, our understanding of specific function of individual geranylgeranyl diphosphate synthases (GGPPS) paralogs is limited (Ament et al., 2006; Jassbi et al., 2008; Schmidt et al., 2010). Reports on basic characterization of individual GGPPS isozymes from *A. thaliana* date back more than a decade ago (Zhu et al., 1997a,b; Okada et al., 2000), being completed only in the recent years (Wang and Dixon, 2009; Beck et al., 2013). This emphasizes the difficulties of studying multiple paralog gene families *in vivo*.

According to our current knowledge, 10 GGPPS (GGPPS1-GGPPS4 and GGPPS6-GGPPS11) out of 12 predicted paralogs from *A. thaliana* are functional, i.e., GGPP is the major product they synthesize *in vitro* and/or they complement *E. coli* strains engineered to synthesize lycopene but lacking GGPPS activity (Zhu et al., 1997a,b; Okada et al., 2000; Wang and Dixon, 2009; Beck et al., 2013).

Furthermore, the GGPPSs from *A. thaliana* reside in distinct subcellular compartments and have distinct expression patterns during plant development. GGPPS1 is targeted to mitochondria, GGPPS3 and GGPPS4 to the ER, GGPPS2 and GGPPS6-GGPPS11 to plastids (Zhu et al., 1997a,b; Okada et al., 2000; Wang and Dixon, 2009; Beck et al., 2013). *GGPPS11* is ubiquitously and abundantly expressed, mainly in photosynthetically active tissues (Okada et al., 2000; Beck et al., 2013), likely providing the GGPP substrate for biosynthesis of essential photosynthesis-related isoprenoid compounds such as chlorophylls, carotenoids, phylloquinones or plastoquinones. *GGPPS1-GGPPS10* expression is different during plant development. These paralogs are expressed predominantly in specific root or seed tissues (Beck et al., 2013). Additionally, *GGPPS5* was proposed to be a pseudogene based on sequence analysis (Beck et al., 2013), whereas GGPPS12, the most distant paralog from all predicted GGPP synthases in *A. thaliana*, does not have GGPP synthase activity (Okada et al., 2000; Wang and Dixon, 2009; Beck et al., 2013). However, GGPPS12 seems to be active as a heterodimer and together with GGPPS11 can synthesize geranyl diphosphate (GPP) (Wang and Dixon, 2009).

The localization in different subcellular compartments as well as the distinct expression pattern suggest specific roles for the GGPPS paralogs during *A. thaliana* development. Yet, the biological significance of a highly expanded GGPP branch point and the relationship between the sequence and function of the GGPPS isozymes is not fully understood.

Here, we investigate the evolutionary relationships and molecular characteristics of the GGPPS homologs in plants using a combination of computational analyses and integration with meta-analysis of existing data sets. We identified the GGPPS homologs from 48 plant species representing major plant lineages (green algae, mosses, gymnosperms, and angiosperms) and inferred their evolutionary relationships. We show that multiple within-species GGPPS paralogs exist in several land plants lineages, particularly in angiosperms. The presence of GGPPS paralogs in the moss *P. patens* suggests that GGPPS duplicated early after the diversification from green algae. In gymnosperms, molecular changes in the GGPPS protein domain may have enabled the transition from biosynthesis of primary GGPP-derived compounds to specialized GPP (geranyl diphosphate) metabolites, which play roles in plant-environment interactions. In land plants, a lineage-specific expansion trend of GGPPS is observed.

We have particularly focused on the model plant *A. thaliana* whose nuclear genome retained 12 GGPPS (Lange and Ghassemian, 2003), the highest number of GGPPS paralogs in plants whose genomes have been sequenced to date. Our results suggest that the expansion of the GGPPS family in *A. thaliana* occurred at distinct time points in evolution and by different duplication mechanisms. *GGPPS12*, *GGPPS2-4*, and *GGPPS11* diverged first. *GGPPS2-4* and *GGPPS11* arose during the most recent WGD event that occurred in *A. thaliana*. In contrast, the most recently diverged paralogs (*GGPPS6*, *GGPPS7*, *GGPPS9*, and *GGPPS10*) arose by tandem and segmental genome duplication. Moreover, we hypothesized that if the GGPPS paralogs from *A. thaliana* are not redundant, their persistence in the genome might be attributed to acquired neo- or subfunctionalization. To test this hypothesis, we have inferred the expression states of individual *GGPPS* during plant development. Subsequently, we have mapped these expression states onto the phylogenetic tree of the GGPPS paralogs from *A. thaliana* and inferred the most parsimonious expression pattern of the ancestral GGPPS gene. A statistically significant correlation of sequence and expression divergence substantiated our hypothesis of subfunctionalization in terms of differential expression pattern.

MATERIALS AND METHODS

SEQUENCE RETRIEVAL AND PHYLOGENETIC ANALYSIS

To study the phylogeny of the GGPPS family a rooted maximum-likelihood (ML) tree from 119 homologous protein sequences spanning 48 plant genomes was reconstructed as follows. First, the homologs were selected by searching sequences (i.e., protein sequences including targeting peptides) similar to the 12 predicted GGPPS proteins from *A. thaliana* in the UniProtKB database (The UniProt Consortium, 2009) augmented with the *A. lyrata* genome retrieved from Ensembl Plants v3 (Kersey et al., 2010). The current protein model for GGPPS5 reposited at TAIR v.10 (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp), which proposes that the translation could be initiated at an alternative start codon, resulting in a protein that lacks a plastidial targeting sequence at the N terminus but has a conserved polyprenyl synthase domain was used (Beck et al., 2013).

To qualify as a homolog, sequences had to exceed a Dayhoff alignment score of 130 to all GGPPS from *A. thaliana* proteins using Darwin's Align function (Gonnet et al., 2000). From this set of homologs, a multiple sequence alignment (MSA) was reconstructed (**Supplementary Dataset 1**) using the Mafft FFT-NS-2 method (Katoh and Toh, 2008). From the resulting MSA, a maximum likelihood tree was reconstructed using the PhyML 3.0 software (Guindon and Gascuel, 2003; Guindon et al., 2009). The default parameters were kept, i.e., we have used the *LG* amino acid substitution matrices (Le and Gascuel, 2008), without invariant sites and with four discrete rate categories chosen according to an estimated gamma shape parameter. The reconstruction was done 50 times from different starting topologies and the overall highest scoring reconstruction was kept for the subsequent analysis. Branch support values were computed using the approximate likelihood ratio test (aLRT) (Anisimova and Gascuel, 2006). To root the phylogenetic tree, a parsimony-based method was used (Berglund-Sonnhammer et al., 2006). In brief, from all possible rootings the tree which minimized the number of implied duplication events and gene losses was chosen. Finally, to infer internal nodes of the tree as speciation or duplication nodes we used the species overlap method, which does not assume a particular species phylogeny (Van Der Heijden et al., 2007). In brief, at every inner node of the gene tree, the overlap of species that are present in each of the two subtrees were counted. In cases one species appeared on both sides of the gene tree, a duplication was inferred; else a speciation event was inferred.

Relative divergence dates of the *GGPPS* paralogs from the Arabidopsis lineage were estimated using Bayesian phylogeny reconstruction with the BEAST 1.6.1 and the BEAGLE software (Drummond et al., 2006). From the previously computed MSA, taxa outside the relevant Arabidopsis lineage were removed and the syntenic orthologs from *Carica papaya* were included (CP00020G01300 and CP00158G00190; PGDD database, http://chibba.agtec.uga.edu/duplication/). The aligned amino acid sequences were mapped to their corresponding codon sequences. Using the $ECM + F + \omega + 2_K$ codon substitution model (Kosiol et al., 2007) in the BEAST software, proposition trees for the tree sampling process were generated by a Yule speciation process using an uncorrelated relaxed clock model with logNormal distribution (Drummond et al., 2006). To calibrate the evolutionary timescale, the following normal distribution priors from the literature on the age of two evolutionary events were used: the *A. thaliana* and *A. lyrata* split was set to 13 ± 3 mya (Beilstein et al., 2010) and the stem lineage subtending the eudicot crown group was set to 130 ± 5.5 mya (Davies et al., 2004). The Markov Chain Monte Carlo (MCMC) chain-length was set to 8×10^6. The first 1% of the trees was discarded as burn-in. The TreeAnnotator module from the BEAST software was used to create the consensus trees.

EXPRESSION ANALYSIS

The expression profile map of the GGPPS paralogs from *A. thaliana* was assembled based on ATH1 22K Affymetrix GeneChip microarray data generated by the AtGenExpress Consortium (http://www.weigelworld.org/resources/microarray/AtGenExpress). The AtGenExpress normalized datasets "tissue extended plus" was retrieved from the Bio-Array Resource website (BAR, www.bar.utoronto.ca). Only experiments using wild-type plants were considered. The probesets for the majority of the GGPPS paralogs are specific to their corresponding transcript, except for *GGPPS6* and *GGPPS7* whose transcripts are ambiguously recognized by the same probeset (258121_s_at) due to their high nucleotide sequence similarity. The common expression profiles for these two genes will be referred in figures with the notation "*GGPPS6/7*." Expression values below a threshold of 2.5 (log2 scale) were considered as not detectable on the microarray (Schmid et al., 2005; Beck et al., 2013). Hierarchical agglomerative clustering with a threshold set at a tree height $h = 0.35$ (equivalent to a Pearson correlation coefficient of 0.65) was used to estimate the number of clusters and their composition. The cluster analysis was conducted in R (R Development Core Team, 2010).

ANCESTRAL STATE RECONSTRUCTION AND STATISTICAL ANALYSIS

The ancestral state reconstruction and random permutations were performed with the Mesquite system for phylogenetic computing version 2.75 (Maddison and Maddison, 2011). The character matrix was generated by discretizing the expression clusters, i.e., each expression cluster is assigned to a distinct character state. The ancestral state reconstruction was performed under a parsimony model assuming an unordered model in which all state changes are weighed equally. To evaluate the statistical significance of an observed parsimony score, the data were randomly permuted by reshuffling the discrete states among taxa 1×10^4 times and calculating the parsimony score for each repetition. The *p*-value was estimated from the distribution of the random parsimony scores, as the fraction of random scores (including the observed score) less than or equal to the observed score: $p = (1 + k)/n$ where k is the number of replications with less or as many steps than the actual observed data and n is the total number of replications. A significant phylogenetic signal was observed at a *p*-value smaller than 0.05 (Faith and Cranston, 1991; Wahlberg, 2001).

RESULTS AND DISCUSSION

THE NUMBER OF *GGPPS* GENE PARALOGS INCREASES DURING THE EVOLUTION OF PLANT FUNCTIONAL COMPLEXITY

We have investigated the phylogenetic relationships among GGPPSs from plants to infer evolutionary mechanisms leading to the formation and maintenance of multiple gene copies particularly within the *A. thaliana* genome, which had retained the highest number of paralogs (twelve).

In total, 119 homologous protein sequences exceeding a Dayhoff alignment score of 130 to all GGPPS from *A. thaliana* (see Materials and Methods) were identified and selected for the phylogenetic tree reconstruction. The selected GGPPS homologs represent 48 plant genomes ranging from green algae and mosses to gymnosperms and angiosperms (**Supplementary Table 1**).

The GGPPS phylogenetic tree revealed five main subfamilies, referred here to as sub. I to sub.V (**Figure 1**). Plant-specific *GGPPS* genes might have originated from an ancestral copy that was present in the common ancestor of land plants and green algae. This is in agreement with earlier publications proposing that all trans-isoprenyl diphosphate synthases, an enzyme class including the GGPPSs, are derived from a common ancestral gene whose precise identity as archaeal or bacterial homolog is not fully elucidated to date (Chen et al., 1994; Tachibana et al.,

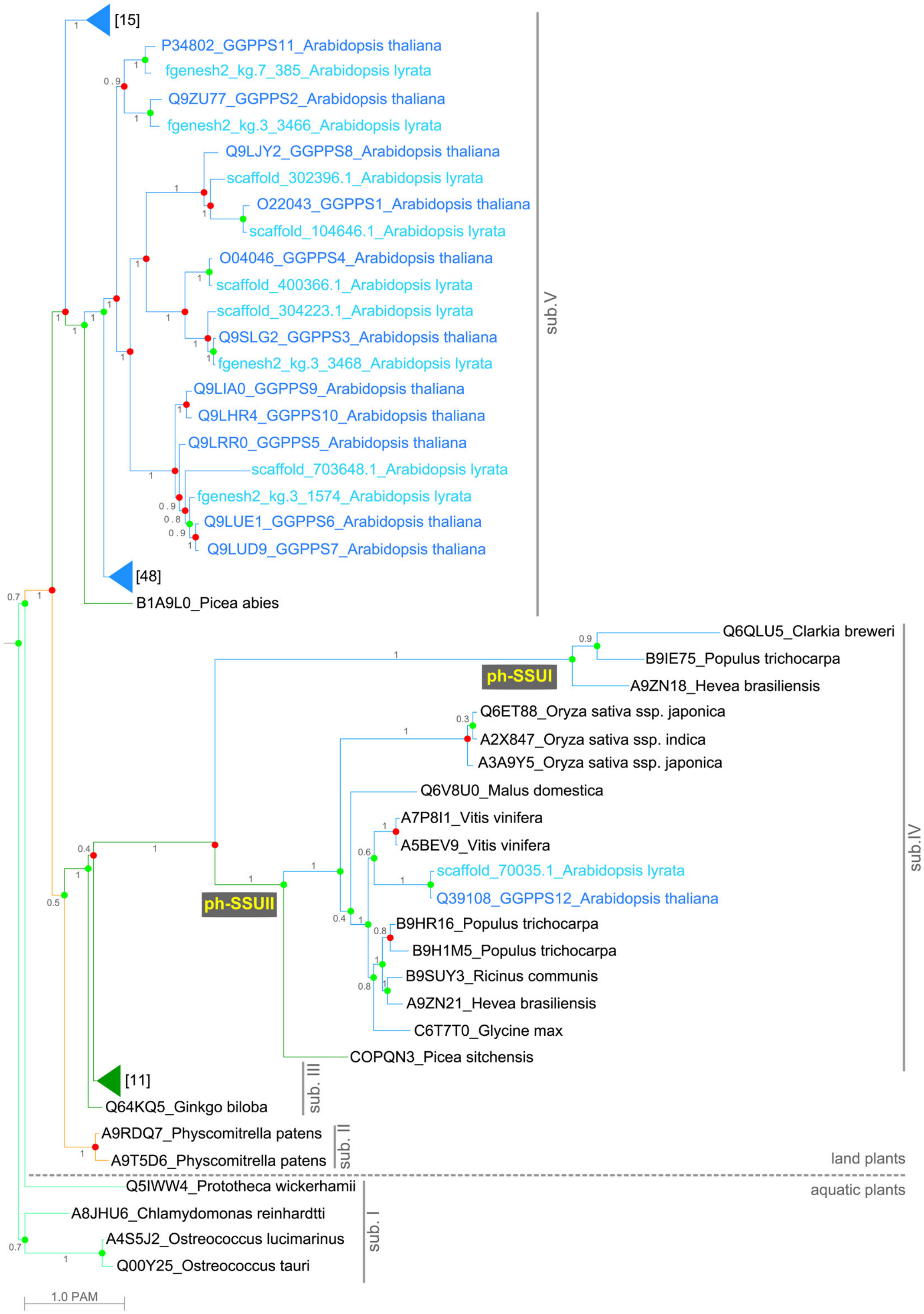
[15]
P34802_GGPPS11_Arabidopsis thaliana
fgenesh2_kg.7_385_Arabidopsis lyrata
Q9ZU77_GGPPS2_Arabidopsis thaliana
fgenesh2_kg.3_3466_Arabidopsis lyrata
Q9LJY2_GGPPS8_Arabidopsis thaliana
scaffold_302396.1_Arabidopsis lyrata
O22043_GGPPS1_Arabidopsis thaliana
scaffold_104646.1_Arabidopsis lyrata
O04046_GGPPS4_Arabidopsis thaliana
scaffold_400366.1_Arabidopsis lyrata
scaffold_304223.1_Arabidopsis lyrata
Q9SLG2_GGPPS3_Arabidopsis thaliana
fgenesh2_kg.3_3468_Arabidopsis lyrata
Q9LIA0_GGPPS9_Arabidopsis thaliana
Q9LHR4_GGPPS10_Arabidopsis thaliana
Q9LRR0_GGPPS5_Arabidopsis thaliana
scaffold_703648.1_Arabidopsis lyrata
fgenesh2_kg.3_1574_Arabidopsis lyrata
Q9LUE1_GGPPS6_Arabidopsis thaliana
Q9LUD9_GGPPS7_Arabidopsis thaliana
[48]
B1A9L0_Picea abies
sub.V
Q6QLU5_Clarkia breweri
B9IE75_Populus trichocarpa
A9ZN18_Hevea brasiliensis
ph-SSUI
Q6ET88_Oryza sativa ssp. japonica
A2X847_Oryza sativa ssp. indica
A3A9Y5_Oryza sativa ssp. japonica
Q6V8U0_Malus domestica
A7P8I1_Vitis vinifera
A5BEV9_Vitis vinifera
scaffold_70035.1_Arabidopsis lyrata
Q39108_GGPPS12_Arabidopsis thaliana
ph-SSUII
B9HR16_Populus trichocarpa
B9H1M5_Populus trichocarpa
B9SUY3_Ricinus communis
A9ZN21_Hevea brasiliensis
C6T7T0_Glycine max
COPQN3_Picea sitchensis
sub.IV
[11]
Q64KQ5_Ginkgo biloba
sub. III
A9RDQ7_Physcomitrella patens
A9T5D6_Physcomitrella patens
sub. II
land plants
Q5IWW4_Prototheca wickerhamii
aquatic plants
A8JHU6_Chlamydomonas reinhardtti
A4S5J2_Ostreococcus lucimarinus
Q00Y25_Ostreococcus tauri
sub. I
1.0 PAM
0.9
0.8
0.7
0.6
0.5
0.4
0.3
1

FIGURE 1 | Maximum likelihood consensus tree of the 119 GGPPS homologs from plants. Posterior probabilities are shown next to the branches. Branch lengths correspond to evolution distances (see Materials and Methods). Duplication (red dots) and speciation (green dots) events are shown at nodes. The tree is divided into five classes (sub. I–V). Branch colors represent the major plant lineages: spring green, green algae; orange, mosses; dark green, gymnosperms; and blue- angiosperms. Branches holding homologs from gymnosperms and angiosperms are collapsed and the number of homologs in each collapsed group is shown. The homologs from the Arabidopsis lineage are shown: in blue-*A. thaliana*, in cyan-*A. lyrata*.

2000). Early after the diversification of land plants, the number of *GGPPS* paralogs per species increases and already in the moss *P. patens* GGPPS appears to be encoded by multiple gene paralogs. Furthermore, the phylogenetic analysis showed lineage-specific expansion and divergence events occurring in land plants (**Figure 1** and **Supplementary Figure 1**). The increase in the predicted number of GGPPSs per species mirrors the increase in complexity of the species. From one GGPPS in green algae (sub. I), three in mosses (sub. II and sub. V) and one to four in gymnosperms (sub. III–V), the number of GGPPS paralogs per species reaches a maximum of twelve copies within angiosperms in *A. thaliana* (sub. V; **Supplementary Table 1**).

THE MOLECULAR EVOLUTION OF THE POLYPRENYL SYNTHASE DOMAIN ENABLES THE NEOFUNCTIONALIZATION OF GGPPS

To gain further insights in molecular changes underlying the evolution of the GGPPS homologs in plants, we have analyzed the evolution of the characteristic polyprenyl synthase domain (Liang et al., 2002). The GGPPS polyprenyl synthase domain has a first aspartate rich motif, FARM (DDxxxxD; x is any amino acid) and a second aspartate rich motif, SARM (DDxxD; x is any amino acid), which are involved in IPP and DMAPP substrate binding and are critical for GGPP biosynthesis (Liang et al., 2002).

Whereas GGPPSs are typically active as homodimers (Vandermoten et al., 2009), heterodimeric complexes between functional GGPPS and SSUI and SSUII (heterodimeric GPP synthase small subunit I and II, respectively) synthesizing GPP have been reported (Burke et al., 1999; Tholl et al., 2004; Wang and Dixon, 2009). SSUI lost both aspartate rich motifs but has two conserved CxxxC motifs (where x is any hydrophobic amino acid) (Tholl et al., 2004). SSUII has conserved FARM and two CxxxC motifs (Burke et al., 1999; Wang and Dixon, 2009). In heterodimeric complexes between functional GGPPS and SSUII, the CxxxC motifs were shown to be important for physical interaction between subunits. Furthermore, such complexes were shown to be able to produce, with increased efficiency, GPP (Wang and Dixon, 2009). GPP can be also produced by homodimeric GPS (geranyl diphosphate synthase) (Hsiao et al., 2008; Schmidt and Gershenzon, 2008). Interestingly, a protein from *A. thaliana* initially classified as GPS (At2g34630; (Bouvier et al., 2000; Van Schie et al., 2007)), which lost the CxxxC motifs but has conserved FARM and SARM, was shown to produce medium (C25) to long (C45) chain isoprenoid products, and was therefore renamed as polyprenyl pyrophosphate synthase (AtPPPS; Hsieh et al., 2011).

The GGPPS homologs from sub. I, II and V have highly conserved FARM, SARM and one CxxxC motif (**Figure 2** and **Supplementary Figure 2**). Homologs from *A. thaliana* with such protein domain structure were shown to be active as homodimers and produce GGPP (Okada et al., 2000; Wang and Dixon, 2009; Beck et al., 2013).

Several homologs from sub. V, have lost the CxxxC motif (**Figure 2**). Such proteins, referred here to as ph-PPPS (putative homologs of polyprenyl pyrophosphate synthase) retain solely FARM and SARM motifs and are found at $d = 7.03$ distance from root supporting their rapid divergence (**Supplementary Figure 1** and **Supplementary Table 2**). The polyprenyl pyrophosphate synthase (AtPPPS, At2g34630) from *A. thaliana*, which can synthesize medium (C25) to long (C45) chain isoprenoid products, has a similar domain structure as the ph-PPPS proteins (Hsieh et al., 2011).

Within sub. III that is found exclusively in gymnosperms, in addition to the conserved FARM and SARM, a prototype of a second CxxxC motif (CxxxS) appears to have been acquired in a common ancestor of Ginkgo, Taxus, Abies and Picea species (**Figure 2**, **Supplementary Figure 1** and **Supplementary Table 2**). A protein with similar domain structure was recently reported to be bifunctional, producing both GPP and GGPP (Schmidt et al., 2010). GPP is the precursor for biosynthesis of monoterpenoids, a class of specialized metabolites which play roles in pollination, seed dispersal and defense mechanisms (Bohlmann and Croteau, 1999). This suggests that the molecular changes in the protein domains of orthologs found in this class may have enabled the transition from biosynthesis of primary GGPP-derived compounds to specialized GPP-derived metabolites. In Abies and Picea species, mutation of the serine residue to cysteine resulted in a conserved second CxxxC motif (**Figure 2**, **Supplementary Figure 1** and **Supplementary Table 2**). The homolog B1A9K6 from *Picea abies* (**Supplementary Table 2**), which retains two conserved CxxxC concomitant with FARM and SARM, was shown to produce only GPP (Schmidt and Gershenzon, 2008).

The GGPPS homologs from sub. IV appear to have experienced faster sequence divergence compared to sub. III, indicated by the branch length (**Figure 1**). Both FARM and SARM are either missing or SARM is mutated in sub. IV but both CxxxC motifs are present (**Figure 2**). Sub. IV comprises of GGPPS from monocots and dicots and one homolog from gymnosperms, most of them being uncharacterized to date (**Figure 1**). Sub. IV is further comprised of two subclasses referred to here as ph-SSUI and ph-SSUII, i.e., putative homologs of the small subunit (SSU) of heterodimeric GPS (Tholl et al., 2004; Wang and Dixon, 2009). Members of both ph-SSUI and ph-SSUII were shown to be active not as GGPPS but as SSU in heterodimeric GPS complexes, producing the GPP (Tholl et al., 2004; Wang and Dixon, 2009). Interestingly, ph-SSUI members are mainly found in flowering plant species (**Figure 2** and **Supplementary Table 2**). They have lost both aspartate rich motifs (**Figure 2**), likely rendering them inactive as homodimeric enzymes. Consistently, the Q6QLU5 homolog from *Clarkia breweri* (**Figure 1**; ph-SSUI) does not produce GGPP (Tholl et al., 2004). A homolog from *Antirrhinum majus*, with similar protein domain structure was shown to form heterodimeric GPS complexes with functional GGPPS and synthesize GPP as main product in reproductive organs (Tholl et al., 2004). In summary, this subclass of proteins with the unique motif organization (lacking both SARM and FARM but retaining both CxxxC motifs) seems to be responsible for monoterpenoids precursor biosynthesis in reproductive plant organs. Members of the ph-SSUII branch from sub. IV have intact FARM but mutated SARM (**Figures 1**, **2** and **Supplementary Table 2**). The GGPPS12 homolog from *A. thaliana* has such a protein domain structure and consequently, is unable to produce GGPP (Okada et al., 2000). Furthermore, similarly to characterized proteins from ph-SSUI (Wang and Dixon, 2009), GGPPS12 forms heterodimeric

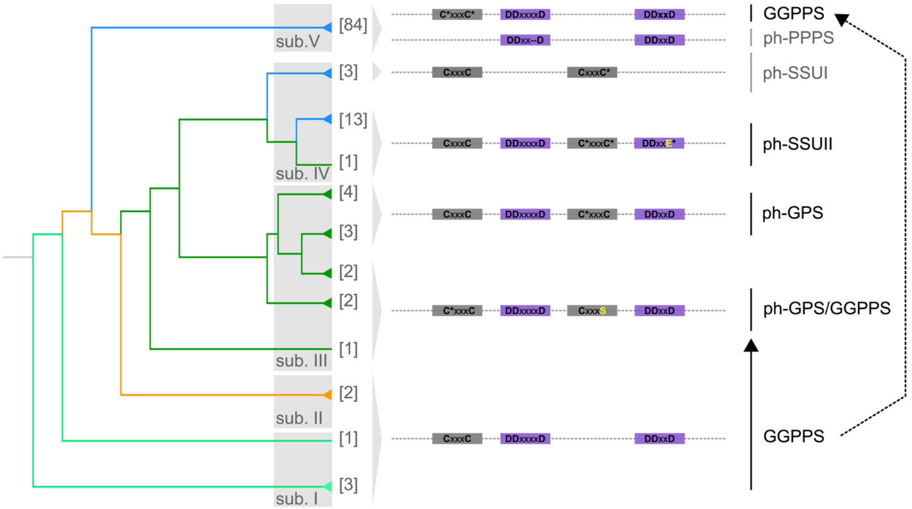

FIGURE 2 | Molecular evolution of the polyprenyl synthase domain. The summarized phylogenetic tree of GGPPS from plants is shown. Branches holding more than one homolog are collapsed and the number of homologs is shown. The five classes (sub. I–V) of GGPPS homologs in plants are shown. Branch colors represent the major plant lineages: spring green, green algae; orange, mosses; dark green, gymnosperms; and blue, angiosperms. The representative polyprenyl synthase motifs for each of the five classes are shown: the two CxxxC motifs in gray and FARM, SARM in purple. Asterisk (*) indicates variable amino acid residues (**Supplementary Figure 1** and **Supplementary Table 2**). ph-GPS: putative homologs of GPS, ph-GPS/GGPPS: putative homologs of the bifunctional GPS/GGPPS. A prototype of the second CxxxC motif (CxxxS; the serine residue is shown in yellow) appears to have been acquired in a common ancestor of gymnosperms. ph-PPPS: putative homologs of polyprenyl pyrophosphate synthase. ph-SSUI and ph-SSUII: putative homologs of the small subunit (SSU) of heterodimeric GPS. Ph-SSUI proteins have lost the two conserved FARM and SARM motifs. None of the ph-SSUII proteins have a conserved SARM (the variable mutated amino acid residue is shown in yellow) indicating loss of GGPPS capacity.

complexes with GGPPS11 and redirects biosynthesis toward GPP (Okada et al., 2000; Wang and Dixon, 2009). In contrast to ph-SSUI homologs, which are likely to play a role in monoterpenoid biosynthesis mainly in reproductive organs, members of the ph-SSUII were proposed, based on their expression pattern, to constitutively participate in GPP biosynthesis during plant development (Wang and Dixon, 2009).

Taken together, GGPPS homologs with canonical protein domain structure are present in all major plant lineages investigated here. Early after the diversification of land plants, duplication events led to multiple GGPPS genes per species, providing raw material for evolutionary change. Yet, with the divergence of land plants their functional complexity and need for defense strategies also diversified.

By neofunctionalization of GGPPS, novel heterodimeric GPS complex formation capacity, and thereby the GPP biosynthesis was enabled by the acquisition of a second CxxxC motif that likely occurred in the ancestor of gymnosperms. GPP serves as precursor of monoterpenes, which are involved in direct defense mechanisms against herbivores or pathogens, they can indirectly protect plants by attracting predators of attacking herbivores, or they can be emitted from floral tissues to attract pollinators (Pichersky and Gershenzon, 2002; Chen et al., 2003; Keeling and Bohlmann, 2006). Members of the ph-PPPS (sub. V), whose protein domains are similar to the AtPPPS from *A. thaliana* (Bouvier et al., 2000; Hsieh et al., 2011) are likely another example of neofunctionalization. They have lost the two CxxxC motifs and in *A. thaliana,* this enzyme is able to generate multiple products with medium to long chain lengths (C25–C45) (Hsieh et al., 2011).

LINEAGE-SPECIFIC EXPANSION OF *GGPPS* IS MOST EVIDENT IN ARABIDOPSIS

Duplication events leading to lineage-specific expansion of GGPPS (i.e., no discernible ortholog in closely related species) occurred in land plants (**Supplementary Figure 1**). The most prominent example of lineage-specific expansion, with respect to our taxon sampling, is found in the Arabidopsis lineage where, the high GGPPSs sequence similarity determines their clustering in the phylogenetic tree (**Figure 1**). The majority of the GGPPS paralogs in *A. thaliana* and its closest relative *A. lyrata* are found in the same clade and are more similar to each other than to homologs from other species, which is supported by the high branch support values (aLRT $\geq$ 0.8). In particular, *A. thaliana* encodes the largest number of paralogs from the species investigated here, including a unique set of GGPPSs (GGPPS6, GGPPS7, GGPPS9, and GGPPS10) found only in this species (**Figure 1**).

Lineage-specific expansion followed by subfunctionalization is known to be an important mechanism for diversification of gene function (Lespinet et al., 2002; Nowick and Stubbs, 2010). For example, the expression of lineage-specific genes in *A. thaliana* was observed to be confined to fewer tissues, where they are involved particularly in abiotic stress responses (Donoghue et al., 2011).

The expression of the *GGPPS* paralogs specific to *A. thaliana* is under strict developmental control, being expressed in specific tissues and at distinct time during plant development (Beck et al., 2013). For example, *GGPPS6* is expressed only in the meristematic zone of the root tip (columella and lateral root cap), whereas *GGPPS10* expression is distributed over the length of the root but not in the root tip (Beck et al., 2013). Together, these indicate that LSG GGPPS paralogs may have special function only at particular stages during plant development and possibly in response to external environmental signals.

SUBFUNCTIONALIZATION MAINTAINS MULTIPLE *GGPPS* PARALOGS IN THE *A. THALIANA* GENOME

Multiple *GGPPS* paralogs might have been maintained in the genome of *A. thaliana* due to the divergence in their expression patterns. There should be no selective constraints blocking this divergence as long as the initial expression pattern of the ancestral gene is maintained. Thus, we expect that the GGPPS paralogs may have specialized functions in *A. thaliana* according to their expression profiles.

To test this hypothesis we mapped *A. thaliana GGPPSs* expression data onto the phylogenetic tree and reconstructed the ancestral expression states (**Figure 3**). Using a comprehensive dataset for gene expression during *A. thaliana* development (see Materials and Methods) we defined eight expression clusters containing the *GGPPS* paralogs referred to as cI-VIII (**Figure 3A**). Next, we mapped the expression clusters as discrete states onto the phylogenetic tree of the *GGPPS* paralogs in *A. thaliana*. The reconstruction of ancestral expression states was performed using the Mesquite v2.75 system for phylogenetic computing (Maddison and Maddison, 2011), which allows the inference of the most likely hypothetical expression states for the ancestral gene under a maximum parsimony model (**Figure 3B**). The expression states (state 1–8) are shown as colored boxes at the terminal branches. A change in color between sister branches indicates a putative divergence in the expression pattern of the paralog.

The ancestral expression pattern, state 2, is represented by an ubiquitous gene expression during plant development (**Figure 3B**). From an evolutionary perspective, ubiquitous expression is characteristic to housekeeping genes, which are generally associated with slower evolutionary rates (Hurst and Smith, 1999; Koonin, 2009). Thus, housekeeping genes are less likely to experience divergence of their expression pattern. As expected, the parsimony reconstruction supports a ubiquitous expression pattern (state 2) of the ancestral *GGPPS* in *A. thaliana* during plant development. *GGPPS11* and *GGPPS12* represent expression state 2, while the expression pattern of the remaining *GGPPS* paralogs appears to be under developmental control. As such, the expression pattern of the *GGPPS* gene family during development diverged during several rounds of duplication. Some of the emerging expression states are clade specific (state 6; **Figure 3B**). However, there is also an example of same or similar expression pattern that appears to have emerged at different positions in the tree. For example, *GGPPS5* and *GGPPS8* are part of the same class V as they have a similar expression pattern ($r = 0.76$) but are found in distinct phylogenetic clades (**Figure 3**). This suggests that these two paralogs may have independently acquired or lost similar cis-regulatory elements responsible for the regulation of expression during development. Furthermore, several paralogs share a similar expression pattern, which likely reflects the short time since their divergence as in the case of *GGPPS9* and *GGPPS10* (**Figure 3B**).

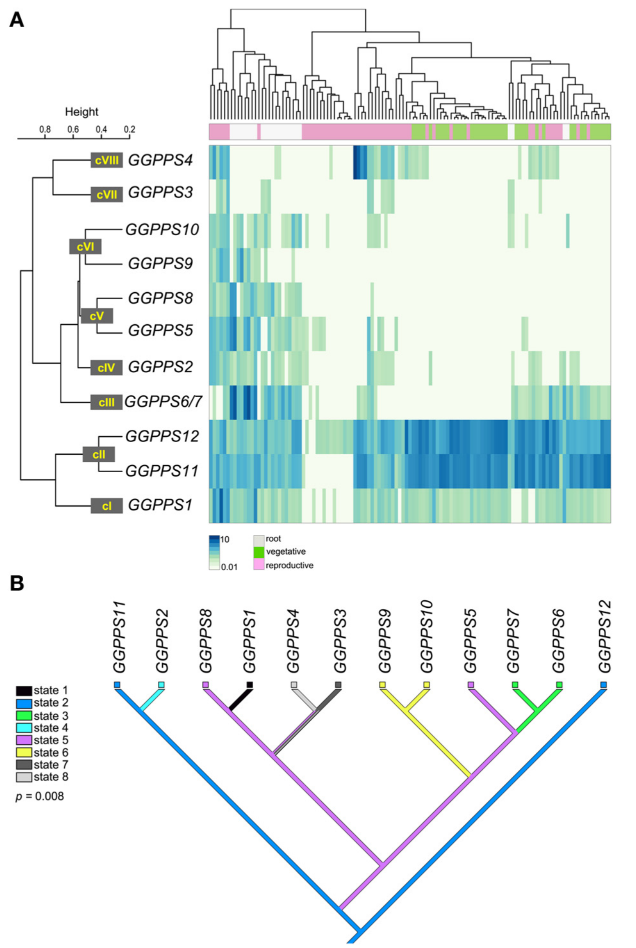

FIGURE 3 | Expression pattern analysis of the *GGPPS* genes from *A. thaliana* and ancestral states reconstruction. (A) The clustering of microarray expression data is shown as heatmap. The expression clusters (cI-VIII) of the *GGPPS* paralogs identified based on Pearson correlation coefficients with a threshold set to $r = 0.65$ (see Materials and Methods) are shown. The various organ and tissue samples were assigned to three major classes: root (white box), vegetative (green box; includes samples from the seedlings, rosette leaves, stems, and cauline leaves) and reproductive (pink box; includes samples from flowers and seeds). **(B)** The phylogenetic reconstruction of ancestral expression states using parsimony is shown. The colors corresponding to each expression state (state 1–8) are shown in the legend. Colored boxes are shown at terminal branches indicating the observed expression pattern cluster. Branches with multiple colors are associated with several possible expression states.

To exclude random events, we evaluated the statistical significance of the correlation between sequence and expression divergence by performing a permutation test in which the expression states were randomly reshuffled. Subsequently, we performed 10,000 ancestral states reconstructions and compared the observed parsimony score against the random distribution from which we calculated the *p*-values. The number of steps required in the random distribution ranged from 7 to 10 in the case of the ancestral states reconstruction of the expression patterns during development. The observed parsimony score of 7 steps indicates non-random distribution that is supported statistically by a permutation *p*-value of 0.008. Therefore, during the evolution of the *GGPPS* gene family in *A. thaliana* the divergence in expression pattern appears to be coupled, at least partially, to sequence divergence.

GGPPS12 and *GGPPS11* genes have an ancestral, ubiquitous expression pattern (**Figure 3**) that may reflect their requirement as housekeeping genes encoding for GGPPS and SSUII, respectively. *GGPPS5* was proposed to encode a pseudogene based on the sequence analysis, which identified a frame shift mutation rendering translation of a truncated GGPPS protein (Beck et al., 2013). Nevertheless, probe based hybridization arrays were able to detect specific expression of *GGPPS5* gene in different organs of *A. thaliana* (**Figure 3**) indicating that *GGPPS5* is an expressed pseudogene also known as ghost pseudogene (Zheng and Gerstein, 2007). As a ghost pseudogene, *GGPPS5* could play a role in regulating the function of closely related paralogs, for example by competing for the cellular RNA degradation machinery (Hirotsune et al., 2003).

GGPPS1 and *GGPPS2* are expressed ubiquitously in all plant organs, but at much lower levels than *GGPPS11* and *GGPPS12* (**Figure 3A**; Beck et al., 2013). *GGPPS3, GGPPS4*, and *GGPPS8* have a mosaic of expression patterns during the plant development. *GGPPS3* and *GGPPS4* are predominantly expressed in reproductive organs and root vasculature, whereas *GGPPS8* is specifically expressed in the outer cell layers above the mitotically active area of the root (**Figure 3A**; Beck et al., 2013). The expression of the *GGPPS* paralogs specific to *A. thaliana* (*GGPPS6*, *GGPPS7*, *GGPPS9*, and *GGPPS10*) is confined to particular tissues (**Figure 3A**; Beck et al., 2013), suggesting that they might play a role only at defined developmental stages and/or in fine tuning adaptation to specific conditions.

Collectively, in addition to neofunctionalization of GGPPS, another mechanism allowing the maintenance of multiple duplicated *GGPPS* paralogs in the *A. thaliana* genome appears to be their subfunctionalization in terms of differential expression pattern during plant development.

THE DUPLICATION TIMING REVEALS A CORRELATION BETWEEN AGE AND EXPRESSION PATTERN OF THE *GGPPSs* FROM *A. THALIANA*

A. thaliana is an ancient polyploid that through evolutionary history experienced three major whole genome duplication events termed γ, β, and α in the order of their occurrence (Bowers et al., 2003). Species such as *Carica papaya* that have not experienced any other whole genome duplication since the γ-WGD event, should have a final set of duplicated genes that have been retained after polyploidisation (Langham et al., 2004; Ming et al., 2008).

To identify the *GGPPS* homologs in *A. thaliana* retained in the *C. papaya* genome, we performed a cross-genome syntenic analysis using the Plant Genome Duplication Database (PGDD, http://chibba.agtec.uga.edu/duplication/). We selected 100 kb of genomic regions adjacent to the *A. thaliana GGPPS* paralogs and the *C. papaya* genome as outgroup. *GGPPS11* and *GGPPS12* are the only paralogs from *A. thaliana*, which have orthologs in syntenic regions of the *C. papaya* genome (**Supplementary Figure 3A**). Next, we have estimated the relative divergence dates of the *GGPPSs* from *A. thaliana*, *A. lyrata* and *C. papaya* based on their codon evolution and using an uncorrelated relaxed clock model (see Materials and Methods).

The molecular-dated phylogenetic tree indicates that after the duplication of an ancestral GGPPS within the time range of the oldest γ-WGD one copy evolved into the common ancestor of the extant *GGPPS12* from *A. thaliana* and its orthologs from *A. lyrata* and *C. papaya*. The other copy duplicated ca. 97 mya and evolved into a *GGPPS* gene in *C. papaya* and into the common ancestor of the remaining 11 extant paralogs in *A. thaliana* (*GGPPS1-GGPPS11*) and their orthologs from *A. lyrata* (**Figure 4**). The *GGPPS* family from the Arabidopsis lineage continued diversifying and expanding during a time range spanning the subsequent β and α-WGD events (**Figure 4**). As such, during the α-WGD, the extant *GGPPS2* and *GGPPS11* arose (ca. 48 mya) followed by *GGPPS3* and *GGPPS4*, which formed ca. 41 mya (**Figure 4**). The remaining extant paralogs (*GGPPS1, GGPPS5*–*GGPPS10*) became fixed in their actual location within the *A. thaliana* genome only after the most recent α-WGD. *GGPPS1* and *GGPPS8* are estimated to have diverged ca. 30 mya, whereas the most recently evolved paralogs in *A. thaliana* are *GGPPS6*, *GGPPS7*, *GGPPS9*, and *GGPPS10*, which arose after sequential duplication of their most recent ancestor between 6 and 9 mya (**Figure 4**).

Generally, following WGD events, many genes return to single copy by fractionation (Lyons et al., 2008). However, some duplicate gene pairs such as genes encoding specialized metabolism enzymes or transcription factors are preferentially maintained (Blanc and Wolfe, 2004; Cannon et al., 2004; Freeling, 2009). Based on the synteny of the surrounding genomic regions, four *GGPPS* paralogs (*GGPPS2*, *GGPPS3*, *GGPPS4*, and *GGPPS11*) are found within α-WGD blocks (Bowers et al., 2003; Thomas et al., 2006) (**Supplementary Figure 3B**). Whereas *GGPPS2* and *GGPPS11* form a pair within one α-WGD block, *GGPPS3* and *GGPPS4* are not retained in pairs with other *GGPPS* paralogs, suggesting that their counterparts were most probably lost due to fractionation processes.

Together, *GGPPS12* appears to be the oldest paralog in *A. thaliana* followed by *GGPPS2-4* and *GGPPS11* (**Figure 4**). Furthermore, *GGPPS2-4* and *GGPPS11* were found in α-WGD blocks and the dated molecular phylogeny confirms their divergences during the time range of the α-WGD, after the ancestor of Arabidopsis split from *C. papaya*. In contrast to the old paralogs in *A. thaliana*, *GGPPS6*, *GGPPS7*, *GGPPS9*, and *GGPPS10* are paralogs specific to *A. thaliana*. After splitting from *A. lyrata*, the genome of *A. thaliana* experienced a 30% reduction in size and at least nine chromosomal rearrangements (Yogeeswaran et al., 2005; Lysak et al., 2006). Thus, it is possible that the *GGPPSs* specific to *A. thaliana* evolved during these genome reshaping events.

The relative age of the *GGPPSs* corresponds to their divergence in their expression pattern. Old paralogs (e.g., *GGPPS11* and *GGPPS12*) are ubiquitously expressed and at high levels whereas young paralogs (e.g., *GGPPS6* and *GGPPS10*) are predominantly expressed in specific tissues and cell types and generally at lower levels (**Figure 3A**; Beck et al., 2013) bringing further indication for subfunctionalization of young paralogs.

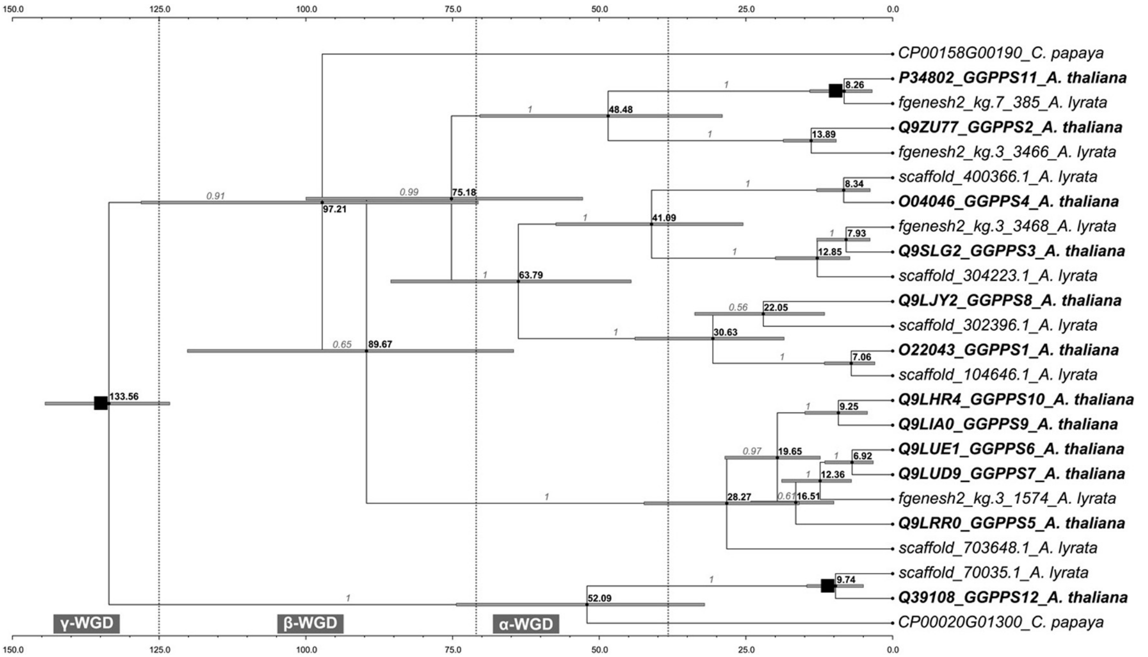
150.0
125.0
100.0
75.0
50.0
25.0
0.0
CP00158G00190_C. papaya
P34802_GGPPS11_A. thaliana
fgenesh2_kg.7_385_A. lyrata
Q9ZU77_GGPPS2_A. thaliana
fgenesh2_kg.3_3466_A. lyrata
scaffold_400366.1_A. lyrata
O04046_GGPPS4_A. thaliana
fgenesh2_kg.3_3468_A. lyrata
Q9SLG2_GGPPS3_A. thaliana
scaffold_304223.1_A. lyrata
Q9LJY2_GGPPS8_A. thaliana
scaffold_302396.1_A. lyrata
O22043_GGPPS1_A. thaliana
scaffold_104646.1_A. lyrata
Q9LHR4_GGPPS10_A. thaliana
Q9LIA0_GGPPS9_A. thaliana
Q9LUE1_GGPPS6_A. thaliana
Q9LUD9_GGPPS7_A. thaliana
fgenesh2_kg.3_1574_A. lyrata
Q9LRR0_GGPPS5_A. thaliana
scaffold_703648.1_A. lyrata
scaffold_70035.1_A. lyrata
Q39108_GGPPS12_A. thaliana
CP00020G01300_C. papaya
8.26
48.48
13.89
8.34
41.09
7.93
12.85
75.18
97.21
63.79
22.05
30.63
89.67
7.06
133.56
9.25
19.65
6.92
12.36
28.27
16.51
9.74
52.09
0.91
0.99
0.65
0.56
0.97
0.61
1
γ-WGD
β-WGD
α-WGD

FIGURE 4 | The calibrated *GGPPS* chronogram. The maximum clade credibility tree and the estimated divergence dates based on total evidence across 24 homologs from *A.thaliana*, *A. lyrata* and *C. papaya* are shown. Branch support values are shown in gray. Note the difference in the relative order between the two clades holding GGPPS2, GGPPS11 and GGPPS5-GGPPS7, GGPPS9, GGPPS10 from **Figure 1**. Both topologies in **Figures 1**, **4** have high support values but are based on different models of evolution that use amino acid and codon sequences, respectively (see Materials and Methods). Mean divergence dates for all nodes are shown in bold black. Gray bars represent the 95% high posterior density credibility interval for node age. Putative intervals for the WGD events are shown. The most ancient event, common to Arabidopsis, Carica, Vitis, and Populus, is the γ-WGD, which separated monocots and eudicot lineages ca. 125–140 mya (Blanc and Wolfe, 2004; Davies et al., 2004; Jaillon et al., 2007). The following more recent WGDs are assumed to have occurred within the *Brassicales*, with the β event having uncertain position after the point of divergence from *Caricaceae* ca. 72 mya (Ming et al., 2008). The most recent α-WGD that occurred ca. 38–70 mya is placed within the *Brassicaceae* (Bowers et al., 2003; Barker et al., 2009) and predates the divergence of *A. thaliana* and *A. lyrata,* which was estimated to have occurred ca. 13 mya (Beilstein et al., 2010). The nodes used as calibration points are indicated by black squares.

CONCLUSIONS

The *A. thaliana GGPPS* gene family is an interesting example of gene evolution involving gene duplication followed by neo- and subfunctionalization as well as pseudogenization. GGPPS homologs with canonical protein domain structure are present in all major plant lineages investigated in this study. Nevertheless, it is possible that neofunctionalization of GGPPS paralogs enabled optimized biosynthesis of primary and specialized metabolites. Furthermore, it was recently proposed that functionality inference for the polyprenyl transferases, should not solely rely on primary sequence due to promiscuity of this class of enzymes (Wallrapp et al., 2013). In the case of the GGPPS family from *A. thaliana*, 10 out of 12 predicted isozymes were shown, using *in vitro* and/or *E. coli* complementation assays, to produce GGPP as major product (see Introduction; Zhu et al., 1997a,b; Okada et al., 2000; Wang and Dixon, 2009; Beck et al., 2013). Still, one cannot exclude that some GGPPS will produce longer polyprenyl diphosphates, thereby providing further means of neofunctionalization.

Our functional divergence analysis suggests that changes in the expression patterns of the *GGPPS* paralogs occurring after gene duplication led to developmental and/or condition specific functional evolution. The ancestral states reconstruction showed a highly non-random distribution of developmental expression patterns in the phylogeny, indicating a significant degree of coupling between sequence and developmental expression divergence. This has prompted us to predict that preserving paralogs with different expression may be of importance for the functional divergence of the *GGPPS* paralogs in *A. thaliana.* Moreover, it was recently proposed that the distinct subcellular localization of the GGPPS paralogs may enable a differential allocation of GGPP precursors to downstream isoprenoid pathways, and as such provide an additional mean of their maintenance in the genome (Beck et al., 2013).

The evolutionary pattern of the *GGPPS* gene family in plants, including variation in paralog number mirroring evolution of plant complexity, lineage-specific expansion, neo- and subfunctionalization is consistent with the idea of GGPPSs as flexible enzymes that might have evolved to support adaptation to various specific conditions. This evolutionary pattern can be recognized in many other gene families, in particular those involved in the specialized metabolism: the cytochrome P450-dependent monooxygenases (P450s) (Bak et al., 2011), glucosidases (Kliebenstein et al., 2005) or the terpene synthase family (Tholl, 2006).

It will be interesting to examine by functional analyses of ggpps single and multiple mutants whether the newly evolved *GGPPS* paralogs in *A. thaliana* are functionally redundant or have indeed specific roles in adaptation to various conditions in a distinct spatial-temporal fashion and in response to specific environmental conditions.

ACKNOWLEDGMENTS

We would like to thank Dr. Katja Bärenfaller for critically reading the manuscript and Dr. Christophe Dessimoz for valuable discussion and suggestions. This work was supported by a grant from ETH Zurich (TH-51 06-1) and the EU FP7 contract 245143 (TiMet).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00230/abstract

Supplementary Figure 1 | Maximum likelihood consensus tree of the GGPPS homologs from plants. Posterior probabilities are shown. Branch lengths correspond to evolutionary distances. Branch colors represent the major plant lineages: spring green, green algae; orange, mosses; dark green, gymnosperms; and blue, angiosperms.

Supplementary Figure 2 | Amino acid MSA of 119 GGPPS homologs from plants. The CxxxC motifs are shown in gray. The FARM and SARM motifs are shown in purple.

Supplementary Figure 3 | Syntenic relationships of *GGPPS* paralogs from *A. thaliana* using *C. papaya* as outgroup. (A) Blocks duplicated by WGD and harboring *GGPPS11* and *GGPPS12* are shown. Their orthologs found in syntenic region of *C. papaya* genome are indicated by red connecting lines. **(B)** *GGPPS2*, *GGPPS3*, *GGPPS4* and *GGPPS11* paralogs from *A. thaliana* found within α-WGD blocks on chromosome 2 and 4, respectively, are shown. Only *GGPPS2* and *GGPPS11* are retained as a pair (connected by red line), whereas the counterparts of *GGPPS3* and *GGPPS4* appear to have been lost from the corresponding syntenic region. Each genomic region spans 100 kb. The *GGPPS* paralogs and their orthologs from *C. papaya* are shown as red arrows. Blue arrows indicate anchor genes and they are connected by blue lines if retain within a WGD block.

Supplementary Table 1 | 119 GGPPS protein sequences used for the phylogenetic tree reconstruction.

Supplementary Table 2 | Polyprenyl synthase domain evolution.

Supplementary Dataset 1 | MAFFT MSA in FASTA format of 119 GGPPS homologs from plants.

REFERENCES

Ament, K., Van Schie, C. C., Bouwmeester, H. J., Haring, M. A., and Schuurink, R. C. (2006). Induction of a leaf specific geranylgeranyl pyrophosphate synthase and emission of (E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene in tomato are dependent on both jasmonic acid and salicylic acid signaling pathways. *Planta* 224, 1197–1208. doi: 10.1007/s00425-006-0301-5

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552. doi: 10.1080/10635150600755453

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Hofer, R., Paquette, S., et al. (2011). Cytochromes p450. *Arabidopsis Book* 9:e0144. doi: 10.1199/tab.0144

Barker, M. S., Vogel, H., and Schranz, M. E. (2009). Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome Biol. Evol.* 1, 391–399. doi: 10.1093/gbe/evp040

Beck, G., Coman, D., Herren, E., Ruiz-Sola, M. Á., Rodríguez-Concepción, M., Gruissem, W., et al. (2013). Characterization of the GGPP synthase gene family in *Arabidopsis thaliana*. *Plant Mol. Biol.* 82, 393–416. doi: 10.1007/s11103-013-0070-z

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18724–18728. doi: 10.1073/pnas.0909766107

Berglund-Sonnhammer, A. C., Steffansson, P., Betts, M. J., and Liberles, D. A. (2006). Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol.* 63, 240–250. doi: 10.1007/s00239-005-0096-1

Blanc, G., and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691. doi: 10.1105/tpc.021410

Bohlmann, J., and Croteau, R. (1999). "Diversity and variability of terpenoid defences in conifers: molecular genetics, biochemistry and evolution of the terpene synthase gene family in grand fir (Abies Grandis)," in *Insect–Plant Interactions and Induced Plant Defence*, eds D. J. Chadwick and J. A. Goode (Chichester: John Wiley and Sons, Ltd.), 132–149.

Bouvier, F., Rahier, A., and Camara, B. (2005). Biogenesis, molecular regulation and function of plant isoprenoids. *Prog. Lipid Res.* 44, 357–429. doi: 10.1016/j.plipres.2005.09.003

Bouvier, F., Suire, C., D'harlingue, A., Backhaus, R. A., and Camara, B. (2000). Molecular cloning of geranyl diphosphate synthase and compartmentation of monoterpene synthesis in plant cells. *Plant J.* 24, 241–252. doi: 10.1046/j.1365-313x.2000.00875.x

Bowers, J. E., Chapman, B. A., Rong, J. K., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438. doi: 10.1038/nature01521

Burke, C. C., Wildung, M. R., and Croteau, R. (1999). Geranyl diphosphate synthase: cloning, expression, and characterization of this prenyltransferase as a heterodimer. *Proc. Natl. Acad. Sci. U.S.A.* 96, 13062–13067. doi: 10.1073/pnas.96.23.13062

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4:10. doi: 10.1186/1471-22 29-4-10

Chen, A. J., Kroon, P. A., and Poulter, C. D. (1994). Isoprenyl diphosphate synthases - protein-sequence comparisons, a phylogenetic tree, and predictions of secondary structure. *Protein Sci.* 3, 600–607.

Chen, F., Tholl, D., D'auria, J. C., Farooq, A., Pichersky, E., and Gershenzon, J. (2003). Biosynthesis and emission of terpenoid volatiles from Arabidopsis flowers. *Plant Cell* 15, 481–494. doi: 10.1105/tpc.007989

Closa, M., Vranová, E., Bortolotti, C., Bigler, L., Arro, M., Ferrer, A., et al. (2010). The *Arabidopsis thaliana* FPP synthase isozymes have overlapping and specific functions in isoprenoid biosynthesis, and complete loss of FPP synthase activity causes early developmental arrest. *Plant J.* 63, 512–525. doi: 10.1111/j.1365-313X.2010.04253.x

Davies, T. J., Barraclough, T. G., Chase, M. W., Soltis, P. S., Soltis, D. E., and Savolainen, V. (2004). Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc. Natl. Acad. Sci. U.S.A.* 101, 1904–1909. doi: 10.1073/pnas.0308127100

Donoghue, M. T. A., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11:47. doi: 10.1186/1471-2148-11-47

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4: e88. doi: 10.1371/journal.pbio.0040088

Faith, D. P., and Cranston, P. S. (1991). Could a cladogram this short have arisen by chance alone - on permutation tests for cladistic structure. *Cladistics* 7, 1–28. doi: 10.1111/j.1096-0031.1991.tb00020.x

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113. doi: 10.2307/2412448

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453. doi: 10.1146/annurev.arplant.043008.092122

Goldstein, J. L., and Brown, M. S. (1990). Regulation of the mevalonate pathway. *Nature* 343, 425–430. doi: 10.1038/343425a0

Gonnet, G. H., Hallett, M. T., Korostensky, C., and Bernardin, L. (2000). Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* 16, 101–103. doi: 10.1093/bioinformatics/16.2.101

Guindon, S., Delsuc, F., Dufayard, J. F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* 537, 113–137. doi: 10.1007/978-1-59745-251-9_6

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., et al. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423, 91–96. doi: 10.1038/nature01535

Hsiao, Y. Y., Jeng, M. F., Tsai, W. C., Chuang, Y. C., Li, C. Y., Wu, T. S., et al. (2008). A novel homodimeric geranyl diphosphate synthase from the orchid *Phalaenopsis bellina* lacking a DD(X)(2-4)D motif. *Plant J.* 55, 719–733. doi: 10.1111/j.1365-313X.2008.03547.x

Hsieh, F. L., Chang, T. H., Ko, T. P., and Wang, A. H. J. (2011). Structure and mechanism of an Arabidopsis medium/long-chain-length prenyl pyrophosphate synthase. *Plant Physiol.* 155, 1079–1090. doi: 10.1104/pp.110.168799

Hurst, L. D., and Smith, N. G. C. (1999). Do essential genes evolve slowly? *Curr. Biol.* 9, 747–750. doi: 10.1016/S0960-9822(99)80334-0

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–U465. doi: 10.1038/nature 06148

Jassbi, A. R., Gase, K., Hettenhausen, C., Schmidt, A., and Baldwin, I. T. (2008). Silencing geranylgeranyl diphosphate synthase in *Nicotiana attenuata* dramatically impairs resistance to tobacco hornworm. *Plant Physiol.* 146, 974–986. doi: 10.1104/pp.107.108811

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* 9, 286–298. doi: 10.1093/bib/bbn013

Keeling, C. I., and Bohlmann, J. (2006). Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytol.* 170, 657–675. doi: 10.1111/j.1469-8137.2006. 01716.x

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., et al. (2010). Ensembl genomes: extending ensembl across the taxonomic space. *Nucleic Acids Res.* 38, D563–D569. doi: 10.1093/nar/gkp871

Kliebenstein, D. J., Kroymann, J., and Mitchell-Olds, T. (2005). The glucosinolate-myrosinase system in an ecological and evolutionary context. *Curr. Opin. Plant Biol.* 8, 264–271. doi: 10.1016/j.pbi.2005.03.002

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucleic Acids Res.* 37, 1011–1034. doi: 10.1093/nar/gkp089

Kosiol, C., Holmes, I., and Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* 24, 1464–1479. doi: 10.1093/molbev/msm064

Lange, B. M., and Ghassemian, M. (2003). Genome organization in *Arabidopsis thaliana*: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol. Biol.* 51, 925–948. doi: 10.1023/a:1023005504702

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945. doi: 10.1534/genetics.166.2.935

Le, S. Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* 25, 1307–1320. doi: 10.1093/molbev/msn067

Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059. doi: 10.1101/gr.174302

Liang, P. H. (2009). Reaction kinetics, catalytic mechanisms, conformational changes, and inhibitor design for prenyltransferases. *Biochemistry* 48, 6562–6570. doi: 10.1021/bi900371p

Liang, P. H., Ko, T. P., and Wang, A. H. J. (2002). Structure, mechanism and function of prenyltransferases. *Eur. J. Biochem.* 269, 3339–3354. doi: 10.1046/j.1432-1033.2002.03014.x

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.54 94.1151

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H. B., et al. (2008). Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 148, 1772–1781. doi: 10.1104/pp.108.124867

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., Mcbreen, K., and Schubert, I. (2006). Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5224–5229. doi: 10.1073/pnas.0510791103

Maddison, W. P., and Maddison, D. R. (2011). *Mesquite: a Modular System for Evolutionary Analysis. Version 2.75*. Available online at: http://mesquiteproject.org

Ming, R., Hou, S. B., Feng, Y., Yu, Q. Y., Dionne-Laporte, A., Saw, J. H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452, U991–U997. doi: 10.1038/nature06856

Nowick, K., and Stubbs, L. (2010). Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief. Funct. Genomics* 9, 65–78. doi: 10.1093/bfgp/elp056

Okada, K., Saito, T., Nakagawa, T., Kawamukai, M., and Kamiya, Y. (2000). Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in Arabidopsis. *Plant Physiol.* 122, 1045–1056. doi: 10.1104/pp.122.4.1045

Pichersky, E., and Gershenzon, J. (2002). The formation and function of plant volatiles: perfumes for pollinator attraction and defense. *Curr. Opin. Plant Biol.* 5, 237–243. doi: 10.1016/S1369-5266(02)00251-0

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rodríguez-Concepción, M., and Boronat, A. (2002). Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics. *Plant Physiol.* 130, 1079–1089. doi: 10.1104/pp.007138

Rohmer, M. (1999). The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat. Prod. Rep.* 16, 565–574. doi: 10.1039/A709175c

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., et al. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37, 501–506. doi: 10.1038/ng1543

Schmidt, A., and Gershenzon, J. (2008). Cloning and characterization of two different types of geranyl diphosphate synthases from Norway spruce (*Picea abies*). *Phytochemistry* 69, 49–57. doi: 10.1016/j.phytochem.2007.06.022

Schmidt, A., Wachtler, B., Temp, U., Krekling, T., Seguin, A., and Gershenzon, J. (2010). A bifunctional geranyl and geranylgeranyl diphosphate synthase is involved in terpene oleoresin formation in *Picea abies*. *Plant Physiol.* 152, 639–655. doi: 10.1104/pp.109.144691

Studer, R. A., and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* 25, 210–216. doi: 10.1016/j.tig.2009.03.004

Tachibana, A., Yano, Y., Otani, S., Nomura, N., Sako, Y., and Taniguchi, M. (2000). Novel prenyltransferase gene encoding farnesylgeranyl diphosphate synthase from a hyperthermophilic archaeon, *Aeropyrum pernix* - Molecular evolution with alteration in product specificity. *Eur. J. Biochem.* 267, 321–328. doi: 10.1046/j.1432-1327.2000.00967.x

The UniProt Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucl. Acids Res.* 37, D169–D174. doi: 10.1093/nar/gkn664

Tholl, D. (2006). Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Curr. Opin. Plant Biol.* 9, 297–304. doi: 10.1016/j.pbi.2006.03.014

Tholl, D., Kish, C. M., Orlova, I., Sherman, D., Gershenzon, J., Pichersky, E., et al. (2004). Formation of monoterpenes in *Antirrhinum majus* and *Clarkia breweri* flowers involves heterodimeric geranyl diphosphate synthases. *Plant Cell* 16, 977–992. doi: 10.1105/Tpc.020156

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946. doi: 10.1101/gr.4708406

Van Der Heijden, R. T. J. M., Snel, B., Van Noort, V., and Huynen, M. A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8:83. doi: 10.1186/1471-2105-8-83

Van Schie, C. C. N., Ament, K., Schmidt, A., Lange, T., Haring, M. A., and Schuurink, R. C. (2007). Geranyl diphosphate synthase is required for biosynthesis of gibberellins. *Plant J.* 52, 752–762. doi: 10.1111/j.1365-313X.2007.03273.x

Vandermoten, S., Haubruge, E., and Cusson, M. (2009). New insights into short-chain prenyltransferases: structural features, evolutionary history and potential for selective inhibition. *Cell. Mol. Life Sci.* 66, 3685–3695. doi: 10.1007/s00018-009-0100-9

Vranová, E., Coman, D., and Gruissem, W. (2012). Structure and dynamics of the isoprenoid pathway network. *Mol. Plant* 5, 318–333. doi: 10.1093/mp/sss015

Vranová, E., Coman, D., and Gruissem, W. (2013). Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.* 64, 665–700. doi: 10.1146/annurev-arplant-050312-120116

Vranová, E., Hirsch-Hoffmann, M., and Gruissem, W. (2011). AtIPD: a curated database of Arabidopsis isoprenoid pathway models and genes for isoprenoid network analysis. *Plant Physiol.* 156, 1655–1660. doi: 10.1104/pp.111.177758

Wahlberg, N. (2001). The phylogenetics and biochemistry of host-plant specialization in Melitaeine butterflies (Lepidoptera: Nymphalidae). *Evolution* 55, 522–537. doi: 10.1554/0014-3820(2001)055[0522:Tpaboh]2.0.Co;2

Wallrapp, F. H., Pan, J. J., Ramamoorthy, G., Almonacid, D. E., Hillerich, B. S., Seidel, R., et al. (2013). Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1196–E1202. doi: 10.1073/pnas.1300632110

Wang, G., and Dixon, R. A. (2009). Heterodimeric geranyl(geranyl)diphosphate synthase from hop (*Humulus lupulus*) and the evolution of monoterpene biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9914–9919. doi: 10.1073/pnas.0904069106

Yogeeswaran, K., Frary, A., York, T. L., Amenta, A., Lesser, A. H., Nasrallah, J. B., et al. (2005). Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of *A-thaliana*. *Genome Res.* 15, 505–515. doi: 10.1101/gr.3436305

Zheng, D. Y., and Gerstein, M. B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* 23, 219–224. doi: 10.1016/j.tig.2007.03.003

Zhu, X. F., Suzuki, K., Okada, K., Tanaka, K., Nakagawa, T., Kawamukai, M., et al. (1997a). Cloning and functional expression of a novel geranylgeranyl pyrophosphate synthase gene from *Arabidopsis thaliana* in *Escherichia coli*. *Plant Cell Physiol.* 38, 357–361.

Zhu, X. F., Suzuki, K., Saito, T., Okada, K., Tanaka, K., Nakagawa, T., et al. (1997b). Geranylgeranyl pyrophosphate synthase encoded by the newly isolated gene GGPS6 from *Arabidopsis thaliana* is localized in mitochondria. *Plant Mol. Biol.* 35, 331–341.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 March 2014; paper pending published: 06 April 2014; accepted: 09 May 2014; published online: 27 May 2014.

Citation: Coman D, Altenhoff A, Zoller S, Gruissem W and Vranová E (2014) Distinct evolutionary strategies in the GGPPS family from plants. Front. Plant Sci. ***5****:230. doi: 10.3389/fpls.2014.00230*

This article was submitted to Plant Evolution and Development, a section of the journal Frontiers in Plant Science.