

Gene Discovery of Modular Diterpene Metabolism in Nonmodel Systems^{1[W][OA]}

Philipp Zerbe², Björn Hamberger², Macaire M.S. Yuen, Angela Chiang, Harpreet K. Sandhu, Lina L. Madilao, Anh Nguyen, Britta Hamberger, Søren Spanner Bach, and Jörg Bohlmann*

Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4 (P.Z., Bj.H., M.M.S.Y., A.C., H.K.S., L.L.M., A.N., Br.H., J.B.); and Department of Plant Biology and Biotechnology, University of Copenhagen, Thorvaldsenvej 40, 1781 Copenhagen, Denmark (Bj.H., Br.H., S.S.B.)

Plants produce over 10,000 different diterpenes of specialized (secondary) metabolism, and fewer diterpenes of general (primary) metabolism. Specialized diterpenes may have functions in ecological interactions of plants with other organisms and also benefit humanity as pharmaceuticals, fragrances, resins, and other industrial bioproducts. Examples of high-value diterpenes are taxol and forskolin pharmaceuticals or ambroxide fragrances. Yields and purity of diterpenes obtained from natural sources or by chemical synthesis are often insufficient for large-volume or high-end applications. Improvement of agricultural or biotechnological diterpene production requires knowledge of biosynthetic genes and enzymes. However, specialized diterpene pathways are extremely diverse across the plant kingdom, and most specialized diterpenes are taxonomically restricted to a few plant species, genera, or families. Consequently, there is no single reference system to guide gene discovery and rapid annotation of specialized diterpene pathways. Functional diversification of genes and plasticity of enzyme functions of these pathways further complicate correct annotation. To address this challenge, we used a set of 10 different plant species to develop a general strategy for diterpene gene discovery in nonmodel systems. The approach combines metabolite-guided transcriptome resources, custom diterpene synthase (diTPS) and cytochrome P450 reference gene databases, phylogenies, and, as shown for select diTPSs, single and coupled enzyme assays using microbial and plant expression systems. In the 10 species, we identified 46 new diTPS candidates and over 400 putatively terpenoid-related P450s in a resource of nearly 1 million predicted transcripts of diterpene-accumulating tissues. Phylogenetic patterns of lineage-specific blooms of genes guided functional characterization.

With more than 10,000 different structures, plant diterpenes are one of the largest and most diverse classes of plant metabolites. All of these compounds are formed from (*E,E,E*)-geranylgeranyl diphosphate (GGPP 1, Fig. 1). A relatively small number of diterpenes of general (i.e. primary) metabolism, such as GAs or chlorophyll (a diterpene conjugate), are known to serve essential functions in plant growth and development. These compounds and their respective biosynthetic pathways are conserved broadly across the

plant kingdom. In contrast, most diterpenes are products of specialized (i.e. secondary) metabolism and may have roles in ecological interactions of plants with other organisms. Examples are antimicrobial diterpene phytoalexins in rice (*Oryza sativa*), wheat (*Triticum aestivum*), maize (*Zea mays*), and other cereal crop plants of the family Poaceae (Peters, 2006; Schmelz et al., 2011; Wu et al., 2012; Zhou et al., 2012a); diterpene resin acids of the oleoresin defense of spruce (*Picea* spp.), pine (*Pinus* spp.), and other species of the Pinaceae family (Keeling and Bohlmann, 2006; Hall et al., 2013); or antiherbivore diterpene glycosides in *Nicotiana attenuata* (Heiling et al., 2010). Individual specialized diterpenes are often of limited taxonomic distribution, where certain metabolites may be found only in a few species, genera, or families. For example, taxol and related taxoids are unique to species of the genus *Taxus* (Jennewein and Croteau, 2001; Guerra-Bubb et al., 2012); labdane-related diterpene phytoalexins are found in several cultivated species of the Poaceae family (Peters, 2006; Schmelz et al., 2011); conifers of the Pinaceae and related families produce an abundance of different diterpene resin acids (Keeling and Bohlmann, 2006; Hamberger et al., 2011; Hall et al., 2013); and different species of wild and cultivated tobacco (*Nicotiana* spp.) produce different profiles of antimicrobial or antiherbivore diterpenes (Heiling et al., 2010; Sallaud et al., 2012; Seo et al., 2012).

¹ This work was supported by the Government of Canada through Genome Canada, Genome British Columbia, Genome Alberta, and Genome Quebec as part of the PhytoMetaSyn Project; by the Natural Sciences and Engineering Research Council of Canada (to J.B.); by the University of British Columbia Distinguished University Scholars program (to J.B.); and by the UNIK Research Initiative of the Danish Ministry of Science, Technology and Innovation and the Novo Nordisk Foundation, through the Center for Synthetic Biology at the University of Copenhagen (to B.H.).

² These authors contributed equally to the article.

* Corresponding author; e-mail bohlmann@msl.ubc.ca.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Jörg Bohlmann (bohlmann@msl.ubc.ca).

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

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

www.plantphysiol.org/cgi/doi/10.1104/pp.113.218347

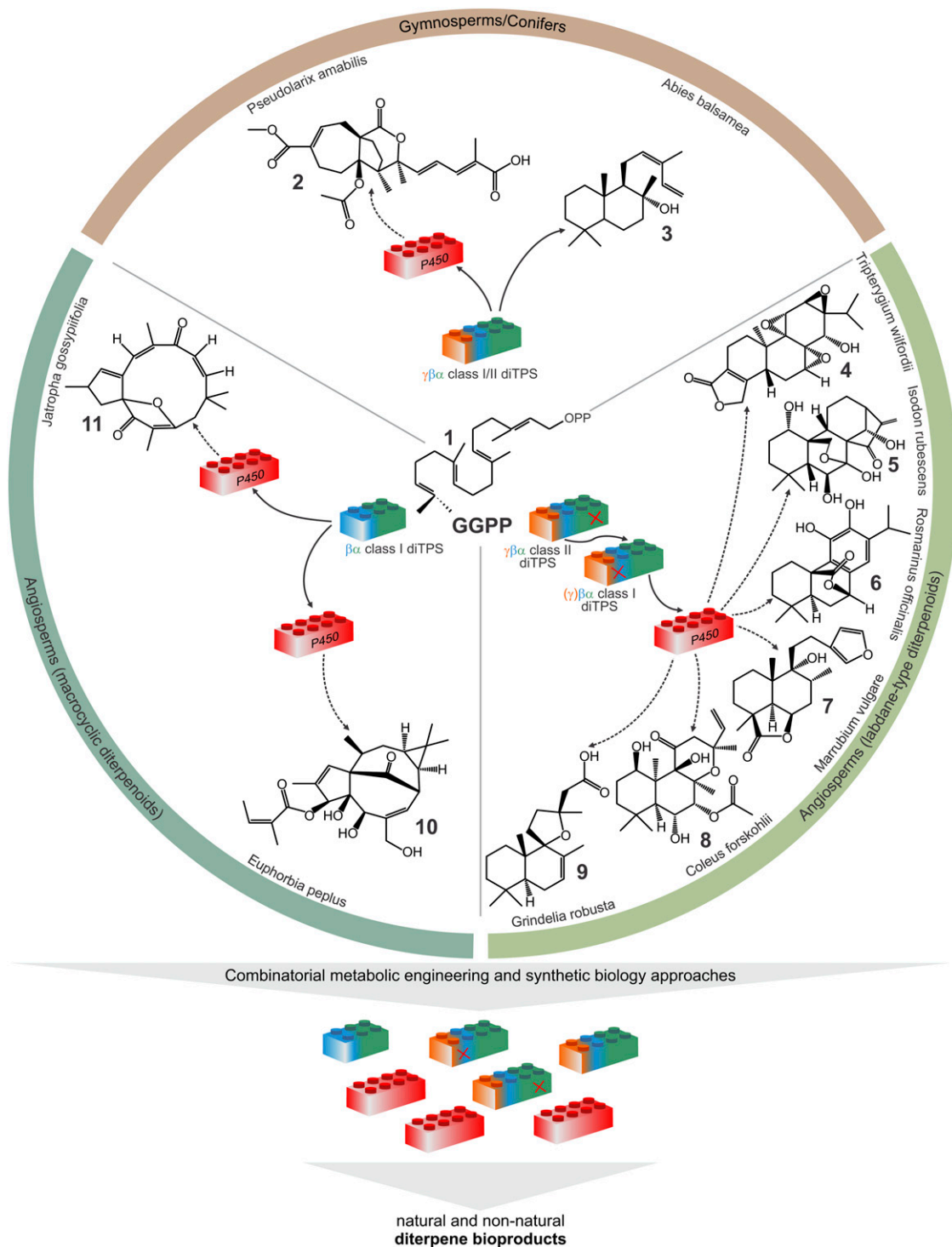


Figure 1. Diterpene biosynthesis. Schematic of proposed pathways leading from GGPP 1 to pseudotsularic acid 2, cis-abienol 3, triptolide 4, oridonin 5, carnosol 6, marrubiin 7, forskolin 8, grindelic acid 9, ingenol-3-angelate 10, and jatrophone 11. These diterpenes are formed by bifunctional class I or class I/II enzymes or pairs of monofunctional class II and class I diTPSs, catalyzing the cycloisomerization of GGPP into distinct diterpene scaffolds. Products of diTPSs can undergo various functional modifications primarily through the activity of P450 enzymes. Considering the modular organization of diterpene specialized metabolism (e.g. Hall et al., 2013), we showed the variable diTPS and P450 enzyme modules as “LEGO-like” blocks. This is conceptually modified from Baldwin (2010), who referred to the five-carbon units of terpenoids as “LEGO” blocks. Here, the different colors of the diTPS modules indicate variations of the $\gamma\beta\alpha$ three-domain structure of plant diTPSs; a red “X” indicates loss of function of class I or class II activity. Recombining diTPS and P450 enzyme modules of diterpene pathway may allow for production of known and new diterpenes through metabolic engineering or synthetic biology.

Plant specialized diterpenes also have a multibillion dollar annual market value as pharmaceuticals, fragrances, resins, and other industrial bioproducts (Bohlmann and Keeling, 2008; Zerbe and Bohlmann, 2013). Diterpene pharmaceuticals include the anticancer drugs taxol derived from *Taxus* species (Jennewein and Croteau, 2001) and ingenol-3-angelate from *Euphorbia peplus* (Li et al., 2010), the cAMP-regulating and vasodilator drug forskolin from *Coleus forskohlii* (Alasbahi and Melzig, 2010b), and the analgesic and antidiabetic drug candidate marrubiin from species of *Marrubium* (Mnonopi et al., 2012). Beyond pharmaceuticals, examples of industrially used plant diterpenes are steviol glycosides sold as natural sweeteners (Goyal et al., 2010), sclareol produced in *Salvia sclarea* (Caniard et al., 2012) as a source of ambroxide fragrances, and diterpene resin acids used as a large feedstock for industrial resins, inks, and coatings (Bohlmann and Keeling, 2008). Sufficient and sustainable access to these compounds can be limited due to lack of cultivation or lack of crop production systems for relevant plant species, low yields from cultivated or noncultivated plant material, overharvesting of wild plant material, or a need for protection of endangered species and habitats. In addition, yields of complex diterpenes and purity of particular stereoisomers obtained by chemical synthesis are often inadequate for large-scale production. Solutions to these issues may be found with improved agricultural or biotechnological production systems. For example, production of taxol uses combinations of precursors isolated from foliage of cultivated *Taxus* species, semisynthesis, and plant cell cultures (Jiang et al., 2012; Wilson and Roberts, 2012). Alternatively, microbial metabolic engineering and synthetic biology approaches are being explored for production of high-value diterpenes (Ajikumar et al., 2010; Caniard et al., 2012; Schalk et al., 2012; Zerbe et al., 2012a; Zhou et al., 2012b). These approaches, however, require knowledge of the specific diterpene pathways, genes, and enzymes.

The structural diversity of specialized diterpenes and their lineage-specific patterns of taxonomic distribution are the result of divergent evolution of specialized diterpene biosynthetic pathways. In addition to sharing a common GGPP 1 starter unit, these pathways are generally built in a “LEGO-like” fashion (compare with Baldwin, 2010) with modules from a few classes of enzymes and along a common building plan (Fig. 1). The common building blocks of modular diterpenoid pathways include enzymes of different classes of diterpene synthases (diTPSs), cytochrome P450-dependent monooxygenases (P450s), and various types of transferases (e.g. acyl-, aryl-, methyl-, or glycosyltransferases) and oxidoreductases. Specific functions of each module may be unique for a given plant species, genus, or family. The modular diterpene pathways can be short, composed for example of only a single diTPS as in the biosynthesis of cis-abienol in *Abies balsamea* (Zerbe et al., 2012a), or employ a suite of up to 20 different modules as in taxol biosynthesis

(Walker and Croteau, 2001). In either case, the committed biosynthetic pathways begin with a diTPS, which in the above examples is either a bifunctional cis-abienol synthase (CAS), which in its particular structure and function is only known from *A. balsamea* (Zerbe et al., 2012a), or a taxadiene synthase (TXS), which is of a unique structure and function that has only ever been found in *Taxus* species (Köksal et al., 2011).

DiTPSs form, through carbocation-driven cyclization reactions, the many different and often polycyclic diterpene scaffolds (Davis and Croteau, 2000; Peters, 2010). Three major classes of plant diTPSs, class I, class II, and class I/II, can be distinguished that differ in their modular architecture of three domains α , β , and γ , and presence of associated catalytic motifs (Köksal et al., 2011; Gao et al., 2012; Zerbe and Bohlmann, 2013; Fig. 1). Class II diTPSs contain a DxDD motif and facilitate the protonation-dependent cyclization of GGPP into distinct bicyclic diphosphate intermediates (e.g. Sun and Kamiya, 1994; Xu et al., 2007a; Falara et al., 2010; Keeling et al., 2010). Class I enzymes contain DDxxD and NSE/DTE signature motifs and catalyze the cleavage of the substrate's diphosphate ester, followed by various possible rearrangements of intermediate carbocations. Known class I diTPSs either convert products of class II diTPSs into bi- or polycyclic diterpenes (e.g. Yamaguchi et al., 1998; Xu et al., 2007a; Caniard et al., 2012; Hall et al., 2013) or convert GGPP into acyclic (Herde et al., 2008; Martin et al., 2010) or macrocyclic diterpene scaffolds (Mau and West, 1994; Hezari et al., 1995; Wildung and Croteau, 1996; Kirby et al., 2010). Bifunctional class I/II diTPSs, which harbor two functional active sites, are only known in nonvascular plants, the lycophyte *Selaginella moellendorffii*, and gymnosperms (e.g. Stofer Vogel et al., 1996; Martin et al., 2004; Hayashi et al., 2006; Mafu et al., 2011; Zerbe et al., 2012a; Hall et al., 2013). Products of the various diTPSs often undergo functionalization commonly starting with the regiospecific oxygenation of the diterpene skeleton through P450 activity, which may be followed by additional modifications that increase the diversity of the metabolite class.

While diterpene biosynthetic pathways found across the plant kingdom employ common modules (classes of enzymes), the functional space within each module is hugely diverse. The catalytic divergence within modules contributes much to the structural diversity of diterpene products and intermediates of these pathways. The diversity of function within modules also has profound implications for gene discovery and gene annotation in the many different species that produce interesting diterpenes. In essence, specific gene and enzyme functions are often unique to specific taxonomic groups. Due to the high level of sequence similarity and functional plasticity within modules, such as the diTPS and P450 modules, specific functions cannot be predicted based on sequence similarity alone. As a consequence, there is no single model or

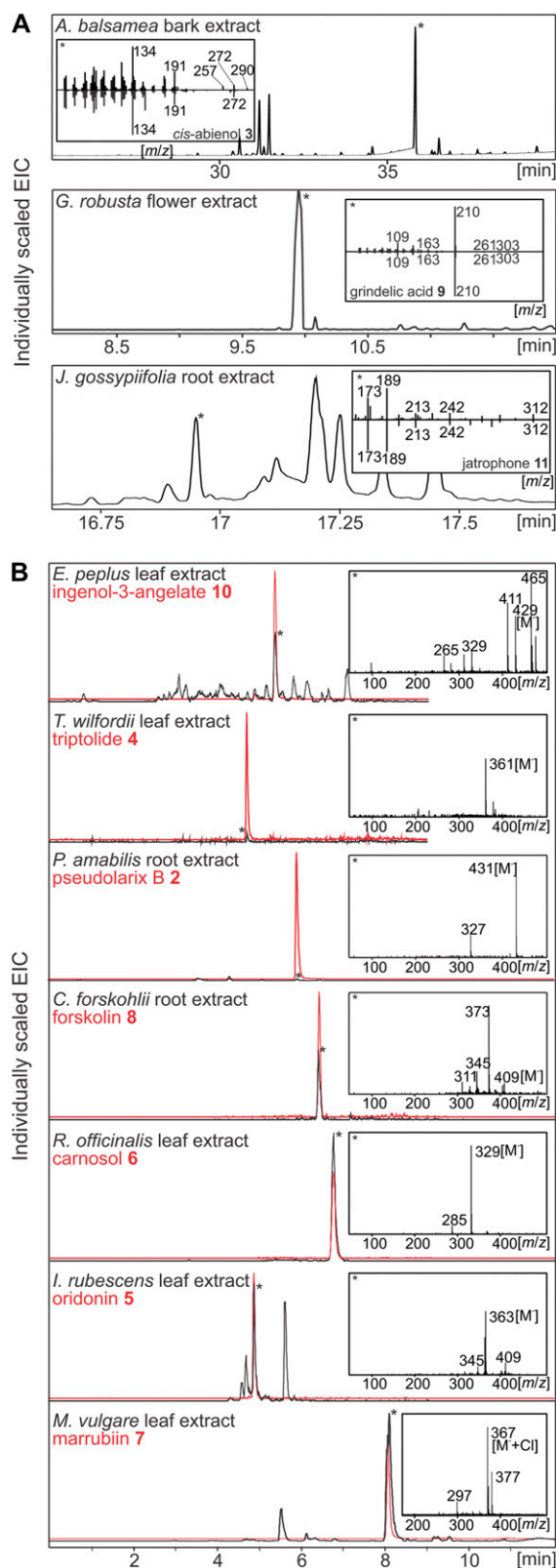


Figure 2. Diterpene profiling. Tissue-specific abundance of diterpene metabolites in the 10 target species of this study was evaluated by GC-MS (A) and LC-MS (B) of tissue extracts with compound verification by comparison to reference mass spectra of the Wiley Registry MS

reference system to guide rapid annotation of the many unique genes and their specific enzymatic functions beyond basic module classifications.

Of the plethora of different plant species that produce ecologically or economically important specialized diterpenes, few have been well characterized for sets of genes and enzymes of the respective biosyntheses. These few include the biosyntheses of taxol in *Taxus* spp. (Walker and Croteau, 2001; Guerra-Bubb et al., 2012), diterpene resin acids in a few conifers (Hamberger et al., 2011; Keeling et al., 2011a; Zerbe et al., 2012a; Hall et al., 2013), and diterpene defense compounds in rice and wheat (Xu et al., 2007a; Wu et al., 2012; Zhou et al., 2012a). In addition, individual enzymes of specialized diterpene biosynthesis, in particular different diTPS genes, have been functionally characterized in approximately a dozen different species. In contrast to the diTPS module, only a few P450s of specialized diterpene pathways have been characterized, specifically P450s of the CYP725 family with functions in taxol biosynthesis (Jennnewein and Croteau, 2001; Jennnewein et al., 2004), P450s of the CYP720B subfamily of diterpene resin acid formation in pine and spruce (Ro et al., 2005; Hamberger et al., 2011), and P450s of the CYP71, CYP99, CYP76, and CYP701 families of diterpene phytoalexin formation in rice (Swaminathan et al., 2009; Wang et al., 2011, 2012a, 2012b; Wu et al., 2011).

Deep or enriched transcriptome resources of diterpene-producing nonmodel systems provide opportunities for the discovery of new genes and enzymes and for exploration of the functional space of specialized diterpene metabolism (Caniard et al., 2012; Facchini et al., 2012; Zerbe et al., 2012a; Higashi and Saito, 2013). It is, however, important to note that accumulation and biosynthesis of specialized diterpenes may be spatially restricted to certain organs, tissue, or cell types or may be temporally limited to certain stages of plant growth and development. In some species, biosynthesis may be induced in response to contact with pathogens or herbivores. Examples are the constitutive and induced production of diterpene resin acids, which in conifers are located in epithelial cells of resin ducts (Zulak and Bohlmann, 2010), the formation of *cis*-abienol in phloem tissue of *A. balsamea* (Zerbe et al., 2012a) or in tobacco leaf trichomes (Ennajdaoui et al., 2010; Sallaud et al., 2012), or the formation of sclareol in flowers of *S. sclarea* (Caniard et al., 2012). Thus, development of transcriptome resources has to be guided by information of metabolite profiles, taking into consideration the fact that biosynthesis and accumulation may be spatially or temporally separated.

Libraries and, where available, authentic standards as detailed in Supplemental Materials and Methods S1. LC-MS chromatograms of tissue extracts (black lines) are compared with authentic standards (red lines), with asterisks indicating the respective diterpenes of interest.

RESULTS

Species Selection and Diterpene Profiling

We selected 10 plant species of five different angiosperm and gymnosperm families for their unique diterpene targets (Fig. 1). We included several species that are being explored, or are already being used, for development or production of diterpene pharmaceuticals: marrubiin 7 produced in *Marrubium vulgare* (Lamiaceae; Meyre-Silva and Cechinel-Filho, 2010), forskolin 8 produced in *Coleus forskohlii* (Lamiaceae; Alasbahi and Melzig, 2010a, 2010b), grindelic acid 9 produced in *Grindelia robusta* (Asteraceae; Zabka et al., 2011), triptolide 4 produced in *Tripterygium wilfordii* (Celastraceae; Brinker et al., 2007), oridonin 5 produced in *Isodon rubescens* (Lamiaceae; Li et al., 2011), pseudolaric acid 2 produced in *Pseudolarix amabilis* (Pinaceae; Chiu et al., 2010), ingenol-3-angelate 10 produced in *Euphorbia peplus* (Euphorbiaceae) (Li et al., 2010), carnosol 6 produced in *Rosmarinus officinalis* (Lamiaceae; López-Jiménez et al., 2011), and jatrophone 11 produced in *Jatropha gossypifolia* (Euphorbiaceae; Theoduloz et al., 2009). In addition, cis-abienol 3 produced in *A. balsamea* (Pinaceae) can be used for production of ambroxide fragrances for high-end perfume manufacture (Zerbe et al., 2012a). For these species we developed diterpene metabolite profiles to identify suitable tissues for transcriptome sequencing. Extracts from roots, stems, and leaves and, where possible, flowers, bark, and wood were analyzed by liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS). Target compounds 2 through 10 were predominantly found in leaves of *E. peplus*, *T. wilfordii*, *R. officinalis*, *I. rubescens*, and *M. vulgare*; in roots of *P. amabilis*, *C. forskohlii*, and *J. gossypifolia*; in bark of *A. balsamea*; and in flowers of *G. robusta*; confirming, in part, earlier reports (Fig. 2).

Transcriptome Sequencing and de Novo Assembly

Nonnormalized complementary DNA (cDNA) libraries were prepared from tissues with highest metabolite abundance for transcriptome sequencing using Roche 454 GS-FLX Titanium and Illumina GAIIx or HiSeq platforms. A one-half-plate 454 sequencing reaction per library yielded a total of approximately 5,500,000 (98% of total reads) high-quality sequences, averaging 610,377 reads per species (Supplemental Table S1), compared with approximately 1,500,000,000 high quality reads (96%), with 59,648,341 and 256,120,403 reads per library on average, obtained with one lane of Illumina GAIIx or HiSeq, respectively. De novo transcriptome assemblies of 454 data sets using a Roche Newbler 2.6 assembler generated 188,270 isogroups (i.e. unique transcripts) with on average 4% of singletons (Supplemental Data S1–S10). The length distribution of isotigs ranged from 76 to 9,530 bp with a median length of 943 bp (Supplemental Fig. S1). Illumina reads were assembled using a Trinity de novo

assembler (Grabherr et al., 2011) and yielded 1,164,906 contigs that built 712,222 components (compare with isogroups) per species (Supplemental Data S1–S10). Contig length varied from 200 to 20,019 bp, showing with 765 bp a shorter median length as compared with 454 assemblies. Mapping of raw Illumina reads against the respective Trinity assemblies showed on average 70% of reads with significant pairing to the assembly (Supplemental Table S1). Overall gene coverage of the transcriptome inventories was validated by comparing all assemblies to the Arabidopsis (*Arabidopsis thaliana*) Core Eukaryotic Genes Mapping Approach (CEGMA) data set (Parra et al., 2007). On average, 92% CEGMA sequences were present in 454 assemblies as compared with 99% in Illumina data sets (defined as query sequence aligning to $\geq 95\%$ in length to top BLASTn hits at $1e^{-20}$ E-value cutoff; Supplemental Table S2), consistent with the typically deeper coverage of Illumina assemblies. Only with *R. officinalis* we obtained lower transcriptome quality with fewer than 74% CEGMA transcripts detected and a higher number of singletons (15%). Of the sequences matching CEGMA targets, 42% of 454 and 57% of Illumina transcripts were full length (FL; defined as 95% coverage compared with top tBLASTn hits at $1e^{-20}$ E-value cutoff). The independent 454 and Illumina assemblies proved useful to validate predicted genes by comparison of the two assemblies for each species.

Targeted Mining of Large Transcriptome Resources for Diterpenoid Biosynthetic Pathways

In plants, the plastidial 2-C-methyl-D-erythritol-4-P (MEP) pathway and the cytosolic mevalonate (MVA) pathway produce five-carbon precursors of isoprenyl diphosphates, with precursors of GGPP produced primarily through the MEP route (Davis and Croteau, 2000). To assess the quality of the transcriptome resources for terpenoid gene discovery, we first evaluated the presence of MEP and MVA genes by mapping the 454 and Illumina assemblies against these two pathways using custom reference sequences from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (E-value threshold $1e^{-20}$). Both pathways, as well as short-chain isoprenyl diphosphate synthases, were completely represented in all 10 species, with 82% of 454-derived and 93% of Illumina-derived transcripts detected as FL sequences (Fig. 3).

Downstream of GGPP biosynthesis, we interrogated the transcriptomes for candidate diTPSs and P450s. To this end, we first developed custom reference databases for plant diTPSs and P450s, which served as a comprehensive suite of baits for transcriptome mining (Supplemental Data S1 and S2). To cover the known variations in terpene synthase (TPS) domain architecture and catalytic functions, the diTPS database comprised a set of class I, class II, and bifunctional class I/II proteins along with select prenyl transferases and mono-, sesqui- and tri-TPSs. Similarly, a representative

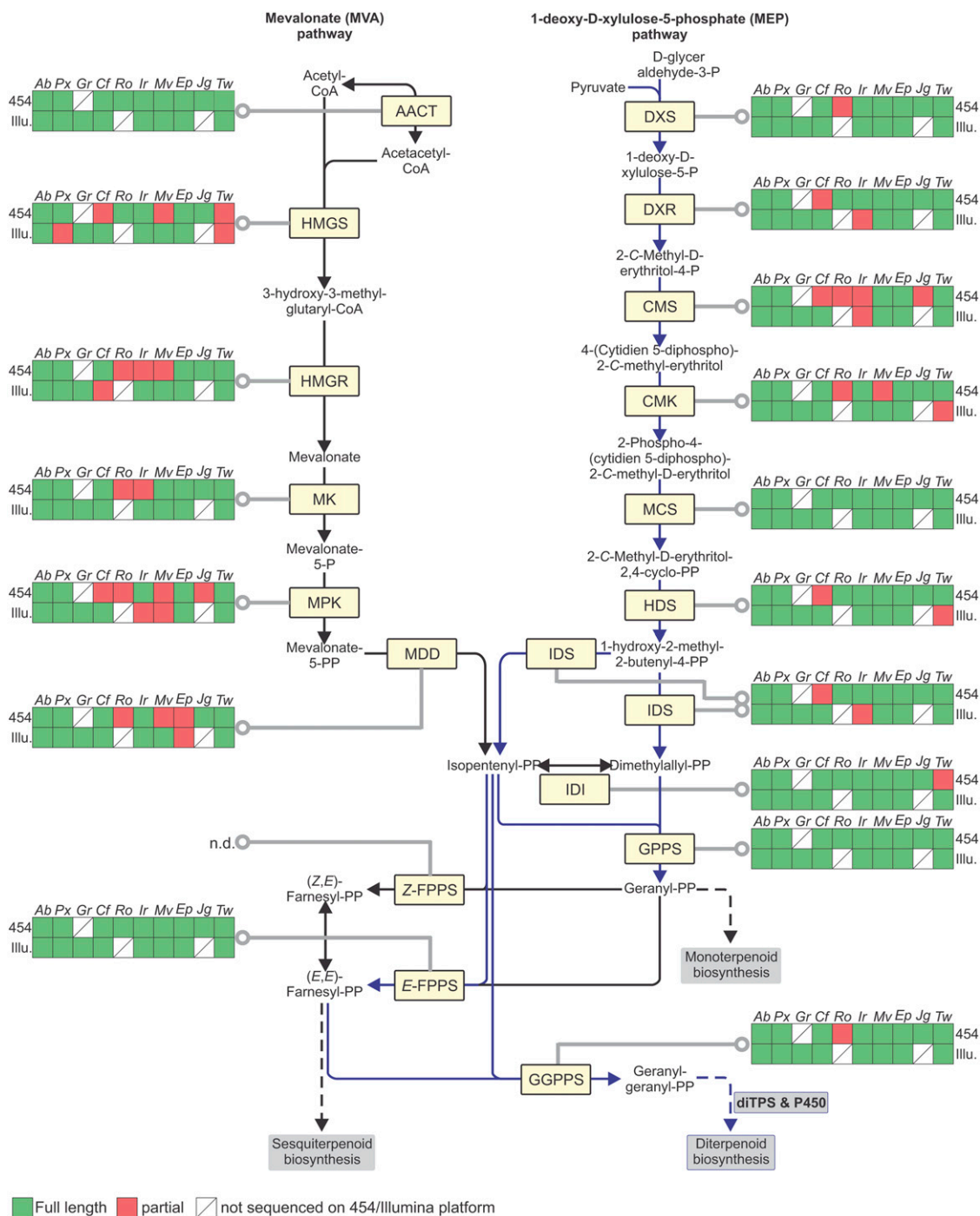


Figure 3. Transcriptome coverage of core terpenoid pathway genes. Coverage of MVA and MEP pathway genes as well as short chain isoprenyl synthases in the transcriptome assemblies was assessed by searches against databases of the KEGG. Blue lines highlight the predominant diterpenoid pathway. AACT, acetyl-CoA-C-acetyl transferase; HMGS, 3-hydroxy-3-methylglutaryl-CoA synthase; HMGR, 3-hydroxy-3-methylglutaryl-CoA reductase; MK, mevalonate kinase; MPK, mevalonate-5-P kinase; MDD, diphospho-mevalonate kinase; DXS, 1-deoxy-D-xylulose-5-P synthase; DXR, 1-deoxy-D-xylulose-5-P reductoisomerase; CMS, 2-C-methyl-erythritol-4-P cytidyl transferase; CMK, 4-(cytidine-5'-diphospho)-2-C-methyl-D-erythritol kinase; MCS, 2-C-methyl-D-erythritol-2,4-cyclo diphosphate synthase; HDS, 1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase; IDS, isopentenyl diphosphate synthase; IDI, isopentenyl diphosphate isomerase; GPPS, geranyl diphosphate synthase; FPPS, farnesyl diphosphate synthase; GGPPS, geranylgeranyl diphosphate synthase.

plant P450 database was designed with at least one representative for each known subfamily. Queries of the transcriptome sequences against these databases identified candidate diTPSs and terpenoid-related P450s, and after removal of redundant transcripts within and across 454 and Illumina assemblies resulted in the identification of 46 new diTPS candidates and more than 400 P450 genes that represent members of terpenoid-related subfamilies (based on an *E*-value threshold of $1e^{-50}$ and ≥ 150 residues in length; Supplemental Fig. S2). Of these, 59% of diTPS candidates and 44% of P450 candidates represented FL sequences. Except for *R. officinalis* (presumably due to the lower assembly quality), all species revealed multimember diTPS and P450 candidate families, consistent with gene family sizes in some related species (Chen et al., 2011; Keeling et al., 2011a; Ma et al., 2012).

Association of diTPS Candidates with Patterns of Evolutionary Divergence of the diTPS Family

The diTPS candidate genes identified in the transcriptomes of the eight angiosperm and two gymnosperm species, together with previously characterized TPSs (Chen et al., 2011), substantiate a phylogeny according to which angiosperm diTPSs of specialized metabolism evolved from ancestral genes along several different branches and in several different subfamilies of the TPS family (Fig. 4).

In the transcriptomes of the two gymnosperm species *P. amabilis* and *A. balsamea*, we identified four diTPSs (*A. balsamea* levopimaradiene/abietadiene synthase [AbLAS], *A. balsamea* isopimaradiene synthase [AbISO], AbCAS, and PxaTPS4) with features of bifunctional class I/II diTPSs of the TPS-d3 group (Fig. 4). Three of these enzymes, AbLAS, AbISO, and AbCAS, were recently functionally characterized (Zerbe et al., 2012a). Bifunctional class I/II diTPSs represent an ancestral form of plant diTPS and have only been found in bryophytes (e.g. *Physcomitrella patens* CPS/KS [for copalyl diphosphate synthase/kaurene synthase]), the vascular nonseed plant *S. moellendorffii*, and gymnosperms (Chen et al., 2011). In *P. amabilis* we also identified three apparently monofunctional $\gamma\beta\alpha$ -domain TPS candidates (PxaTPS1–PxaTPS3). These TPSs contain the DDxxD and NSE/DTE motifs, but lack the DxDD motif, and cluster together with previously characterized gymnosperm bisabolene synthases (Bohlmann et al., 1998b; Martin et al., 2004). This particular class I TPS-d3 group contains diTPS-like sesqui-TPS, which, together with *Taxus* spp. TXS genes, apparently evolved from bifunctional class I/II diTPSs by loss of class II activity and further neofunctionalization (Fig. 4). Evolution of sesqui-TPS functionality required change of substrate specificity (McAndrew et al., 2011). Our phylogeny suggests that loss of class II activity, followed by γ -domain loss further led to the evolution of $\beta\alpha$ -domain sesqui-TPS as well as casbene synthases (CBS) and related diTPSs with macrocyclic products of

the large angiosperm TPS-a subfamily (Fig. 4). Based on the topology of the phylogenetic tree, GGPP-converting macrocyclases appear to have derived from sesqui-TPS through events of reverted evolution with a substrate change from C_{15} to C_{20} .

Species-specific patterns of repeated diTPS gene duplications and sequence divergence (referred to as “blooms”; compare with Feyereisen, 2011) became apparent for several species of the transcriptome analysis, and most prominently with EpTPS and TwTPS, members of the TPS-a subfamily. At larger taxonomic scales, blooms of angiosperm specialized metabolism diTPSs of the large TPS-c and TPS-e/f subfamilies are rooted in *ent*-CPS (ECPS) and *ent*-KS (EKS) genes, respectively, of GA biosynthesis. In contrast, no such blooms are found for gymnosperm diTPSs of the TPS-c and TPS-e/f subfamilies represented in the phylogeny with *Picea sitchensis* ECPS and EKS genes (Fig. 4).

The TPS-c and TPS-e/f subfamilies contain exclusively monofunctional, respectively, class II and class I enzymes, which originated through divergent subfunctionalization from an ancestral bifunctional class I/II diTPS represented in the phylogeny by PpCPS/KS (Fig. 4). Five of the *G. robusta* diTPSs identified in our transcriptome analysis clustered within Asteraceae-specific subgroups of the TPS-c and TPS-e/f subfamilies. Likewise, clusters of Lamiaceae diTPSs were found in the TPS-c and TPS-e/f subfamilies. Potentially of significance for functional annotation, several class II diTPSs from *M. vulgare*, *R. officinalis*, *C. forskohlii*, and *I. rubescens* clustered with characterized labda-13-en-8-ol diphosphate (LPP) synthases from *S. sclarea* (SsLPPS; Caniard et al., 2012) and *Nicotiana tabacum* (NtLPPS; Sallaud et al., 2012), and likewise a (+)-copalyl diphosphate (CPP) synthase from *Salvia miltiorrhiza* (SmCPS; Gao et al., 2009). SsLPPS, NtLPPS, and SmCPS catalyze the class II reactions in coupled reactions with class I diTPSs, leading to the formation of, respectively, sclareol-, cis-abienol-, and tanshinone-specialized diterpenes. Moreover, several *M. vulgare*, *C. forskohlii*, and *I. rubescens* class I enzymes form a perhaps Lamiaceae-specific subgroup of unusual $\beta\alpha$ -domain diTPSs together with *S. sclarea* sclareol synthase (SsSS; Caniard et al., 2012) and *S. miltiorrhiza* miltiradiene synthase (SmMS; Gao et al., 2009), suggesting specialized functionalities (Fig. 4).

Phylogenetic Relationships and Blooms of Candidate P450s of the CYP71 and CYP85 Clans

Known P450s of terpenoid biosynthesis do not represent a monophyletic group, but are found across the huge P450 superfamily in seven of the 10 P450 clans of vascular plants: CYP51, CYP71, CYP72, CYP85, CYP86, CYP710, and CYP711 (Nelson and Werck-Reichhart, 2011; Hamberger and Bak, 2013). Since only very few P450s of diterpene metabolism are functionally characterized, it is currently not possible

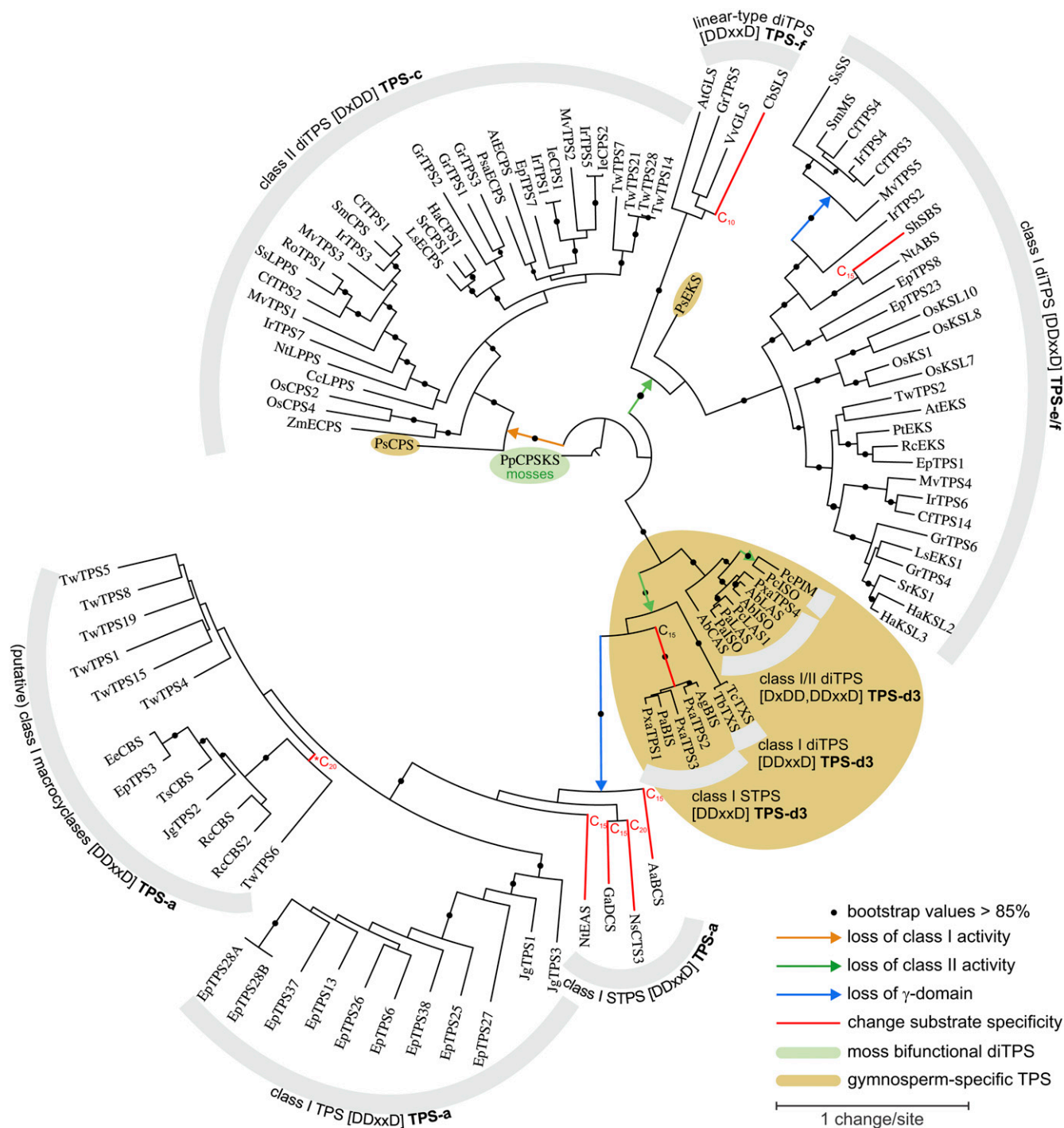


Figure 4. Phylogeny of candidate diTPSs. Maximum likelihood tree, illustrating phylogenetic relationships of diTPS candidates relatively to previously known diTPSs. *P. patens* copalyl diphosphate synthase/kaurene synthase (PpCPS/KS) was used as the tree root. For abbreviations and accession numbers see Supplemental Table S3.

to reconstruct the evolution of P450s of diterpene metabolism. Ancestral P450s of the CYP71 and the CYP85 clans with general diterpene GA biosynthesis and triterpene biosynthesis may have served as progenitors for the evolution of P450s of

specialized terpene metabolism, since these two clans show features of blooms of terpenoid P450s (Hamberger and Bak, 2013). The CYP71 and the CYP85 clans have previously been mined for the discovery of P450s of specialized diterpene biosynthesis, specifically

biosynthesis of taxol (CYP725A subfamily of the CYP85 clan; Jennewein and Croteau, 2001; Jennewein et al., 2004), diterpene resin acids (CYP720B subfamily of the CYP85 clan; Ro et al., 2005; Hamberger et al., 2011), and rice diterpene phytoalexins (CYP99, CYP76 and CYP701 families of the CYP71 clan; Swaminathan et al., 2009; Wang et al., 2011, 2012a, 2012b; Wu et al., 2011).

Querying the transcriptomes of the 10 diterpene specialized metabolite producing species, we found a total of well over 1,000 candidate P450s, of which up to 440 genes fall into terpenoid-related subfamilies of the CYP71 and CYP85 clans. Within these clans we identified expansive lineage-specific blooms of P450 subfamilies as shown in Figure 5 for four of the 10 species. These blooms are indicative of roles in specialized metabolism. In *T. wilfordii*, six members of a novel CYP88H subfamily (CYP85 clan) were found, in addition to the related TwCYP88A (Fig. 5A). While CYP88A genes function in GA biosynthesis and often occur as single copy genes, the TwCYP88H subfamily identified here is distantly related to CYP88D6 from licorice (*Glycyrrhiza glabra*) involved in triterpene biosynthesis of glycyrrhizin (Seki et al., 2008), indicative of roles in specialized metabolism. In *C. forskohlii*, we observed a large expansion of the CYP716 family (CYP85 clan; Fig. 5B). Five *C. forskohlii* CYP716 genes fall into a clade of CfCYP716A subfamily members, two *C. forskohlii* genes each belong to subfamilies CYP716D and CYP716E, and one *C. forskohlii* gene represents the CYP716C subfamily. Based on read counts, CfCYP716C1, CfCYP716A3, and CfCYP716A4 were highly overrepresented in the 454-sequenced transcriptome. In the gymnosperm *P. amabilis*, we found two large clusters of in total 18 members of the CYP76AA subfamily (CYP71 clan) that group closely with *P. sitchensis* CYP76Z1 (Fig. 5C). Lack of angiosperm representatives in this CYP76 branch indicates a gymnosperm-specific family diversification and suggests that angiosperm CYP76 genes evolved from a common ancestor with the apparently gymnosperm-specific CYP76Z and CYP76AA subfamilies. In *E. peplus*, several closely related subfamilies of the CYP71 clan showed a high degree of expansion (Fig. 5D). A total of 23 *E. peplus* CYP71 candidates form a largely expanded CYP71D tribe comprising 16 members and a smaller 726A subfamily with seven candidates. In particular, members of the large CYP71D bloom are of interest, with several CYP71D enzymes being involved in terpenoid biosynthesis across the plant kingdom (Ralston et al., 2001; Wüst et al., 2001; Wang and Wagner, 2003).

Functional Characterization of diTPS Candidates

Definitive functional annotation of genes such as diTPSs and P450s that belong to multigene families with divergent functions and large catalytic plasticity of the encoded enzymes requires experimental verification of predicted functions. While attempting this for the

several hundred P450s identified here is beyond the scope of the present work, we selected a few diTPS candidates to validate the identification of new diTPSs and the feasibility of both in vitro assays with microbial expressed proteins and in vivo assays based on transient (co)expression in *Nicotiana benthamiana* using single diTPSs as well as combinations of diTPSs.

In vitro activity assays showed that PxaTPS4 is a functional LAS-type diTPS, converting GGPP into epimers of the diterpene alcohol 13-hydroxy-8(14)-abietene that dehydrate to form a mixture of abietadiene, palustradiene, levopimaradiene, and neoabietadiene (Fig. 6A). Consistent with its phylogenetic position among other class I/II diTPSs in the TPS-d3 family (Fig. 4), PxaTPS4 is a functional ortholog of LAS enzymes of spruce (*Picea* spp.), fir (*Abies* spp.), and pine (*Pinus* spp.; Stofer Vogel et al., 1996; Keeling et al., 2011b; Zerbe et al., 2012a; Hall et al., 2013).

We tested three different *G. robusta* diTPS candidates of Asteraceae-specific subgroups of the TPS-c and TPS-e/f subfamilies using in vitro and in vivo assays (Figs. 6B–D and 7A). Concurrent with its position next to Arabidopsis and grapevine (*Vitis vinifera*) geranylinalool synthases (GLSs) on a distinct branch of the TPS-f subfamily, GrTPS5 was functionally identified as a GLS by comparison of the enzyme product to authentic geranylinalool (Fig. 6B). GrTPS1 was identified as a class II LPPS by comparison to the product profile of a previously reported protein variant of AbCAS (Fig. 6C; Zerbe et al., 2012a). Consistent with the AbCAS variant and other LPP synthases (LPPSs) (Caniard et al., 2012), GC-MS analysis of the dephosphorylated GrTPS1 product showed, next to labd-13-en-8,15-diol (i.e. dephosphorylated LPP), a racemic mixture of manoyl oxide and *epi*-manoyl oxide (Fig. 6C). The latter diterpenes could not be detected without dephosphorylation of the GrTPS1 product, indicating that they represent derivatives formed under GC-MS conditions. While no in vitro or in vivo activity could be detected for GrTPS6 alone, combination of GrTPS6 with GrTPS1 in in vitro and in vivo assays resulted in formation of manoyl oxide with only traces of *epi*-manoyl oxide (Figs. 6D and 7A), demonstrating a new class I function for GrTPS6 that proceeds via cleavage of the diphosphate group of LPP and subsequent heterocyclic ring closure including the hydroxyl group at C-8.

We also tested the possibility of using combinations of diTPSs from different species. Products with high similarity to pimaradiene and abietadiene were observed, when coupling GrTPS6 with a maize ECPS (Harris et al., 2005) and in trace amounts when coupled with a (+)-CPP-producing variant of Norway spruce (*Picea abies*) LAS (Zerbe et al., 2012a; Fig. 6D).

In agreement with its close phylogenetic relationship to known *Euphorbia esula* and *Triadica sebifera* CBS enzymes of the TPS-a family, heterologous expression of the $\beta\alpha$ -domain EpTPS3 in *N. benthamiana* resulted in the formation of casbene (Fig. 7B) as identified by comparison with reference mass spectra (Reiling et al.,

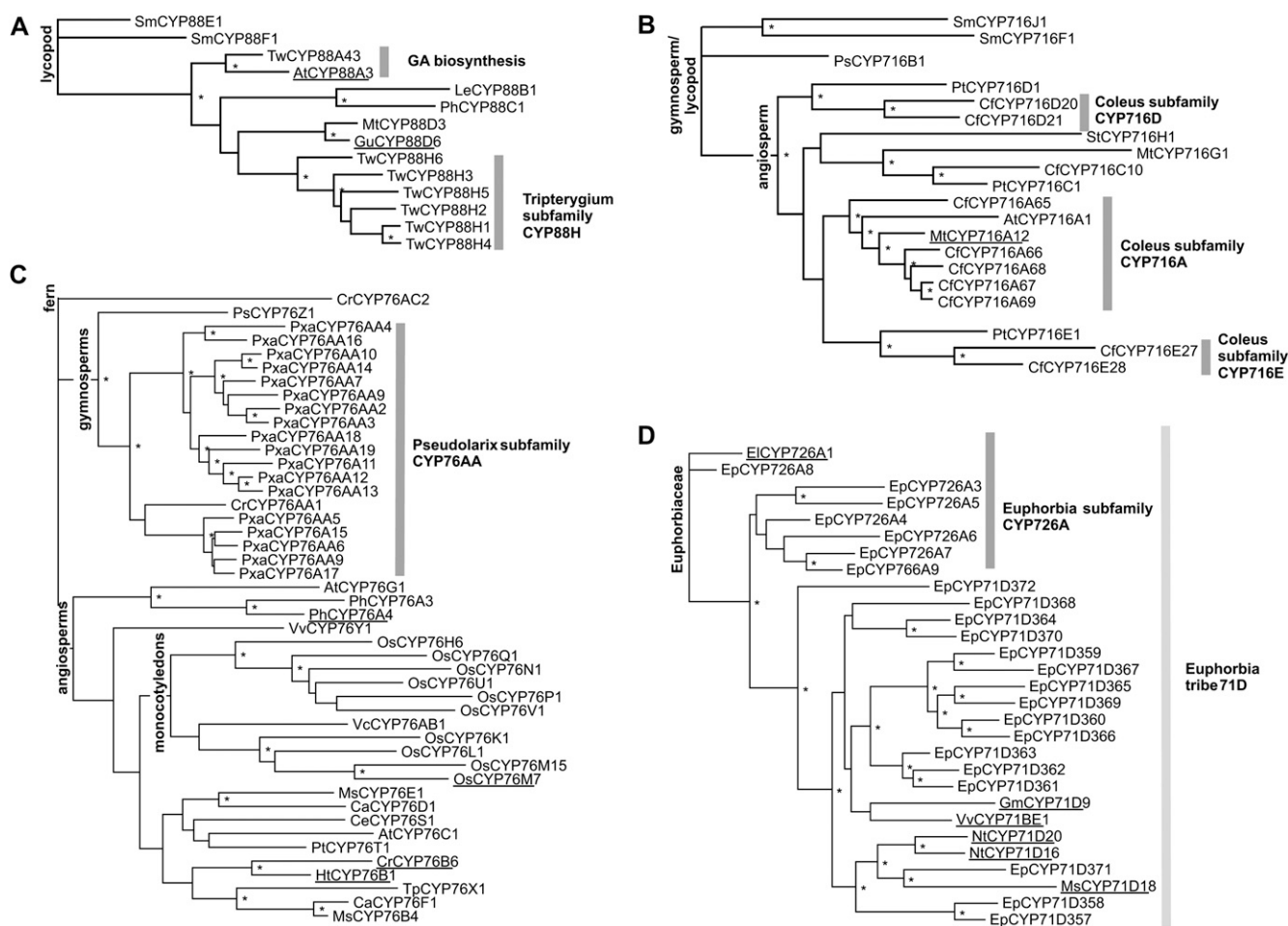


Figure 5. Phylogenetic analysis of P450 candidates of four select species. Maximum likelihood trees of CYP88 members in *T. wilfordii* (A), CYP716 candidates of *C. forskohlii* (B), CYP76 candidates of *P. amabilis* (C), and members of the CYP71D and CYP726A families in *E. peplus* (D). Sequences of previously characterized P450s are underlined. Asterisks indicate bootstrap support of greater than or equal to 80%. Phylogenetic trees were rooted with ancestral representatives. Subfamilies represent select manually curated P450s, including all available subfamilies. For abbreviations and accession numbers see Supplemental Table S4.

2004). *E. peplus* TPS, a member of the class II TPS-c subfamily, and EpTPS1 and *C. forskohlii* TPS14, both belonging to the class I TPS-e/f subfamily, were also analyzed using the *N. benthamiana* in vivo assay system (Fig. 7). While expression of the individual enzymes in *N. benthamiana* yielded no detectable new products, coexpression of EpTPS7 with EpTPS1 or CfTPS14 resulted in accumulation of *ent*-kaurene (Fig. 7C) and suggested functions of these three enzymes in GA biosynthesis.

DISCUSSION

Mining of Plant Biodiversity for Bioproduct Genes and Synthetic Biology

Plants are estimated to produce more than one-quarter million different specialized metabolites (Pichersky and Lewinsohn, 2011), which each typically

require several unique enzymes for biosynthesis. To accomplish this biosynthetic diversity, nature must have evolved and maintained across the biodiversity of plant species hundreds of thousands of more or less closely related enzymes for specialized metabolism. Discovery and accurate annotation of the corresponding genes of specialized metabolism pose a challenge for high-throughput biology, but if successful also present new opportunities for development of biocatalysts, bioproducts, and synthetic biology. Discovering and exploring the enormous enzyme diversity of plant specialized metabolism on a large scale requires genomic and biochemical research beyond traditional model systems. Annotation of the divergent enzymes of specialized metabolism is further supported by the development of databases that accurately capture the functional space of enzymes in specialized metabolism, where minor sequence variations may substantially affect substrate or product specificity.

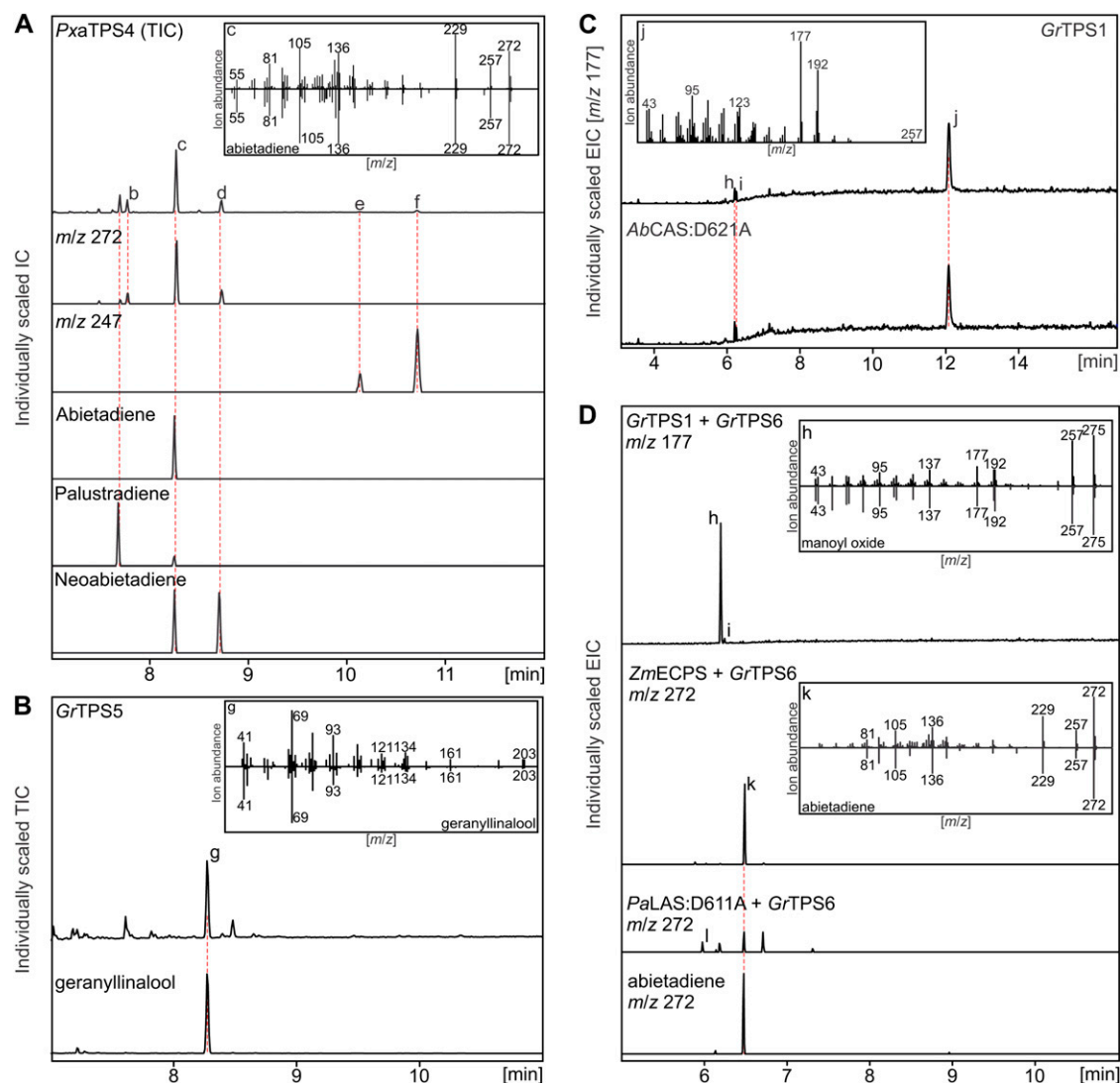


Figure 6. In vitro characterization of recombinant diTPSs from *G. robusta* and *P. amabilis*. GC-MS traces of reaction products from single or coupled in vitro assays of recombinant proteins with 15 μM GGPP 1 as substrate. Major products are compared with authentic standards or reference mass spectra of the Wiley Registry MS Libraries. A, Diterpene alcohol and olefin formation by PxaTPS4. B, Geranylinalool production by GrTPS5. C, Formation of LPP with manoyl oxide and *epi*-manoyl oxide byproducts by GrTPS1 verified through identification of the dephosphorylated reaction product (labd-13-en-8,15-diol). D, Activity of GrTPS6, forming manoyl oxide and traces of *epi*-manoyl oxide in combination with GrTPS1, and abietadiene when coupled with ECPS (ZmECPS; Harris et al., 2005) or (+)-CPS (PaLAS:D611A; Zerbe et al., 2012a). IS, Internal standard 1.6 μM eicosene; peak a, palustradiene; peak b, levopimaradiene; peak c, abietadiene; peak d, neoabietadiene; peak e/f, epimers of 13-hydroxy-8(14)-abietene; peak g, geranylinalool; peak h, manoyl oxide; peak i, *epi*-manoyl oxide; peak j, labd-13-en-8,15-diol; peak k, abietadiene; peak l, pimaradiene-type diterpene. AbCAS:D621A, *A. balsamea* CAS variant, producing LPP (Zerbe et al., 2012a).

In this study we focused on diterpene specialized metabolites for proof-of-concept work in an important and very large class of plant-derived bioproducts. The many roles of plant diterpenes in the natural world and their applications as pharmaceuticals, fragrances, resins, food additives, and other commercial bioproducts have spurred a growing interest in exploring and harnessing the biodiversity of diterpene metabolism. Because diterpene biosynthesis is generally

organized with pathways of common modules (Fig. 1), we began to investigate select modules and the gene space and functional diversity within such modules. We reasoned that this approach can provide a rational starting point for a general strategy of enzyme discovery of diterpenoid biosynthesis that could likewise apply to other classes of terpenes, such as the thousands of monoterpenes and sesquiterpenes. The pathways of mono-, sesqui- and diterpenes differ in the

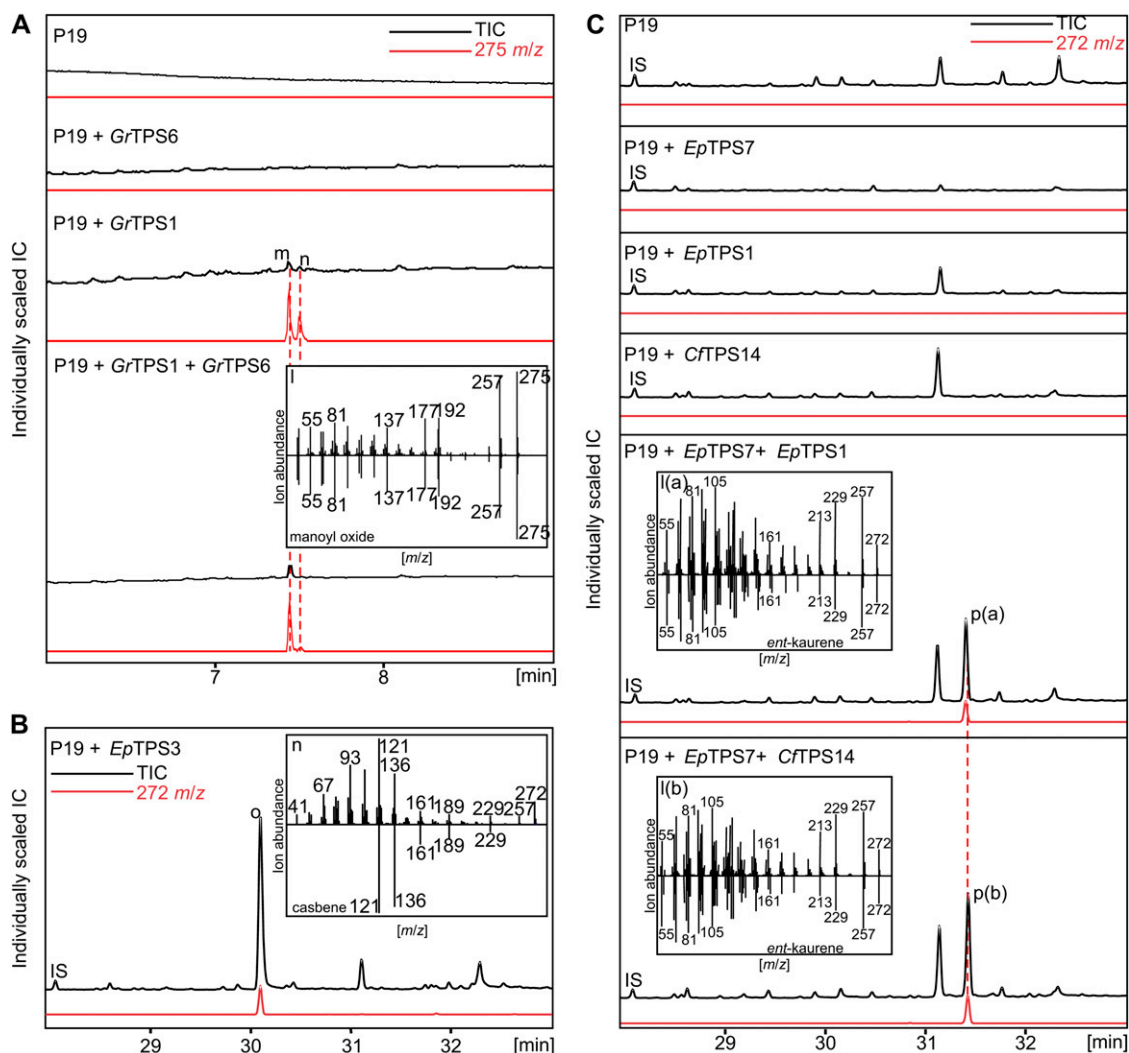


Figure 7. In vivo characterization of diTPSs expressed in *N. benthamiana*. GC-MS traces of *A. tumefaciens*-infiltrated *N. benthamiana* leaf extracts after transient expression of GrTPS1, GrTPS6, and combination of both diTPSs, in the presence of the P19 silencing suppressor strain (A); EpTPS3 (B); and EpTPS1, EpTPS7, CfTPS14, and the combinations EpTPS7/EpTPS1 and EpTPS7/CfTPS14 (C). Transformation with P19 alone served as a control. Compound accumulation was analyzed 4 d post infiltration. Peak m, Manoyl oxide; peak n, *epi*-manoyl oxide; peak o, casbene; peak p(a, b), *ent*-kaurene; IS, internal standard 0.2 mg⁻¹ octadecane.

chain length of a few common isoprenyl diphosphate substrates at the beginning of each pathway, with GGPP 1 being the common precursor for diterpenes. The first diversifying module of terpenoid biosynthesis is the TPS module, which includes a large number of diTPSs, in addition to hemi-TPS, mono-TPS, and sesqui-TPS (Bohlmann et al., 1998a; Chen et al., 2011). In this work, we identified features of the TPS phylogeny that will guide general identification of diTPSs relative to hemi-TPS, mono-TPS, or sesqui-TPS, and may also provide informed direction for exploring specific functions (see below).

Following the diTPS module, the P450 superfamily is a huge reservoir of enzymes that may modify diTPS products and provide pathway end products or

intermediates for further functional decoration. Compared with the diTPSs, much less is known about the P450 module of diterpene biosynthesis, but our work highlights that plant species with diterpene specialized metabolism show blooms of P450s within clans, families, or subfamilies of the P450 superfamily that are promising candidates for terpene metabolism. The functional characterization of any P450 in terpene biosynthesis is a difficult task, often limited by availability of authentic metabolite standards and daunting in the face of the large number of putative candidate genes. Developing in vivo systems for producing relevant diterpene substrates by metabolic engineering of diTPSs and combinations thereof in heterologous hosts provides platforms in which candidate P450 genes can

subsequently be tested. The usefulness of platform diTPS expression systems for P450 discovery has been demonstrated in previous work (Ro et al., 2005; Swaminathan et al., 2009; Hamberger et al., 2011; Wang et al., 2011, 2012a, 2012b; Wu et al., 2011), and we have extended here on the development of diTPS-expressing platforms using both yeast (*Saccharomyces cerevisiae*) and plant expression systems, which will be suitable for additional expression of P450s. Selecting candidate P450s from expanded lineage-specific tribes and subfamilies within relevant clades of the P450 superfamily greatly reduces the number of candidates to be tested in a first screen with in vivo platform expression systems. This approach of selecting candidates from lineage-specific P450 blooms proved successful in the past for the discovery of a new gymnosperm subfamily of CYP720B genes of the Pinaceae family that produce a diverse array of diterpene resin acids (Hamberger et al., 2011), and should now be applicable to other systems based on mining of deep transcriptome resources from diterpene-producing tissues.

Our approach of individual and combinatorial diTPS expression also already moved beyond proof of concept to the identification of diTPS functionalities, which have not previously been described, as illustrated here by the stereospecific formation of manoyl oxide through sequential activity of GrTPS1 (class II) and GrTPS6 (class I). The particular functions of manoyl oxide formation with combined GrTPS1 and GrTPS6 activity can now be utilized for developing diterpene platforms for additional gene discovery and production of hydroxylated diterpene therapeutics, such as forskolin 8. We also showed here the utility to explore substrate promiscuity and new diTPS combinations with GrTPS6, which also converted *ent*-CPP and to a lesser extent (+)-CPP into pimaradiene- and abietadiene-type diterpenes, when combining this class I diTPS with different class II diTPSs of other species.

A Functionally Informed and Informative Phylogeny of Plant diTPS

We showed here that plant diTPSs are not restricted to one or a few subfamilies of the TPS family (Chen et al., 2011), but are found in several major subfamilies, specifically TPS-a, TPS-c, TPS-d, TPS-e/f, and TPS-f. diTPSs are not known in the TPS-b and TPS-g subfamilies. diTPSs have evolved as mono- or bifunctional enzymes of different domain structures. Nevertheless, all plant diTPSs can be traced back to a putative bifunctional three-domain class I/II diTPS ancestor, which may be most closely resembled in extant plants by the CPS/KS enzyme of the moss *P. patens* (Fig. 4). Our analysis of diTPSs of different TPS subfamilies highlights a few vignettes of the TPS phylogeny as it pertains to evolution of diTPS functions. The resulting functionally informed diTPS phylogeny will be useful for directing experimental validation of new candidate

diTPSs as shown with a few select examples in this study.

Originating from an ancestral PpCPS/KS-type bifunctional diTPSs, class I/II diTPSs of the gymnosperm specific TPS-d3 clade are a major relatively basal group of the TPS family. LAS and ISO functionality appears to be a more ancient function of the TPS-d3 group. This interpretation is supported by the functional characterization of PxaTPS4, described here, as a functional ortholog to known LAS (Peters et al., 2000; Martin et al., 2004; Keeling et al., 2011a; Zerbe et al., 2012a; Hall et al., 2013), suggesting that LAS functionality was established prior to the divergence of different genera in the Pinaceae. From this basal function evolved, within the gymnosperm TPS-d3 group, monofunctional class I diTPSs (represented with TXS) and sesqui-TPSs of the bisabolene synthase type. Descendants of the TPS-d3 group also include the many gymnosperm mono-TPSs and sesqui-TPSs of the TPS-d1 and TPS-d2 clades (not shown in Fig. 4; see Keeling et al., 2011a). Also branching out from the gymnosperm TPS-d3 group are the large TPS-a group, which includes angiosperm diTPS and sesqui-TPS functions (Fig. 4). In contrast to the relationship between blooms of the gymnosperm TPS-d subfamily and the angiosperm TPS-a subfamily, blooms of angiosperm diTPSs of TPS-c and TPS-e/f subfamilies are not paralleled by similarly expanded families in the gymnosperms. Blooms of angiosperm diTPSs of TPS-c and TPS-e/f subfamilies appear to have evolved through repeated duplication and neofunctionalization of ancestral ECPS and EKS of GA biosynthetic diTPSs (Peters, 2010). In contrast, the bona fide ECPS and EKS genes of gymnosperm GA biosynthesis, positioned at the base of the TPS-c and TPS-e/f subfamilies respectively, appear to be resistant to diversification and are retained as orphans in “single-copy families” (compare with Keeling et al., 2010; De Smet et al., 2013).

Across the 10 species of our analysis we found blooms of CBS-like diTPSs of the TPS-a subfamily in *E. peplus*, *J. gossypifolia* (Euphorbiaceae), and *T. wilfordii* (Celastraceae) with more than 20 genes (Fig. 4). Based on the relatedness with *Ricinus communis* CBS (Mau and West, 1994) and *E. esula* CBS (Kirby et al., 2010), and considering the macrocyclic diterpene biochemistry of *E. peplus* and *J. gossypifolia* (Fig. 1), these genes are good candidates for discovery of macrocyclases; indeed we verified CBS macrocyclase activity for EpTPS3 (Fig. 7B). The diTPS phylogeny also showed emerging patterns of functional clades within the TPS-c and TPS-e/f subfamilies. Within the TPS-c subfamily, ECPS genes of GA biosynthesis cluster together, and in the TPS-e/f subfamily EKS and other KS-like genes cluster together. The informative nature of clustering was supported with functional characterization of EpTPS7 as an ECPS in the TPS-c subfamily, and functional characterization of EpTPS1 and CfTPS14 as EKS in the TPS-e/f subfamilies (Fig. 7C). A separate functionally informative cluster of specialized metabolism diTPSs also emerged around a scaffold of

three different previously described LPPSs (Falara et al., 2010; Caniard et al., 2012; Sallaud et al., 2012), indicating that new candidate class II diTPS that are situated together with these genes may also encode LPPS or related functions and are therefore primary targets for roles in the biosynthesis of hydroxylated diterpenoids, such as marrubiin 7, forskolin 8, and oridonin 5 in *M. vulgare*, *C. forskohlii*, and *I. rubescens*, respectively. Within the TPS-e/f subfamily, a few candidate diTPS from *M. vulgare*, *C. forskohlii*, and *I. rubescens* form a distinct group of class I diTPSs with functionally characterized SsSS (Canard et al., 2012) and SmMS (Gao et al., 2009). These Lamiaceae genes show an unusual $\beta\alpha$ -domain architecture that can be explained by loss of the γ -domain (Fig. 4), as previously reported for a diTPSs from wheat and *S. miltiorhiza* (Hillwig et al., 2011). This cluster of Lamiaceae diTPSs is likely to be involved in specialized metabolism, given its distinct separation from $\gamma\beta\alpha$ -domain class I enzymes and the high abundance of the corresponding transcripts in the target tissue-specific transcriptomes of *M. vulgare*, *C. forskohlii*, and *I. rubescens*. With *G. robusta* TPS5, we functionally identified the first Asteraceae member of the TPS-f family of GLSs (Fig. 6B) in addition to known GLSs from Arabidopsis, grapevine, and *N. attenuata* (Herde et al., 2008; Martin et al., 2010; Jassbi et al., 2008). This group also contains linalool synthase from *Clarkia breweri* (Dudareva et al., 1996), which may have evolved from a GLS function by a simple change of substrate specificity, but retaining formation of an acyclic product.

Despite the emerging patterns of the diTPS phylogeny, which provides functional information, a priori distinction of diTPSs of general and specialized metabolism within the TPS-c and TPS-e/f clades is still not trivial. This is due, in part, to the fact that similar diTPS functions may have evolved in parallel in separate taxonomic groups and the corresponding genes may cluster by taxonomic rather than functional relatedness. For example, we functionally identified *G. robusta* GrTPS1 as an LPPS (Fig. 6C; Fig. 7A) despite its closer relationship to other diTPSs of the Asteraceae than to LPPSs from other species.

Clades of Terpenoid-Related P450s

The enormous size of the P450 superfamily (Nelson and Werck-Reichhart, 2011; Hamberger and Bak, 2013) with very few functionally characterized P450s of diterpene biosynthesis provides a challenge for a priori gene annotations. However, some general observations may guide selection of candidate P450s from transcriptome or genome sequences. First, the majority of P450 families involved in specialized terpene metabolism belong to the CYP71 and CYP85 clans. Second, emergence of lineage-specific P450 blooms (in the P450 nomenclature system typically named as subfamilies) in plant species or families with unique and diverse diterpene metabolism may be indicative of

specialized P450 functions as was shown with the CYP720B subfamily of diterpene resin-producing gymnosperms of the Pinaceae family (Hamberger et al., 2011). Such blooms can distinguish genes of potentially similar functions in general metabolism, as was also shown with the CYP720B example that was clearly distinct from the CYP701A subfamily of *ent*-kaurene oxidation in GA biosynthesis. Here, we found several species-specific blooms within the CYP71 clan with P450 candidates from *P. amabilis* and *E. peplus*, which are now being explored as candidates in the biosynthesis of pseudolaric acid 2 and ingenol-3-angelate 10 (Fig. 5). Similarly, blooms within the CYP85 clan were observed with a diversified CYP88H family from *T. wilfordii* and a bloom of highly abundant members of the CYP716 subfamily in *C. forskohlii* that may include candidates with possible functions in triptolide 4 and forskolin 8 biosynthesis, respectively.

General Applications and Conclusion

We developed an approach that identified nearly 500 diTPS and P450 candidate genes from 10 plant species of five different gymnosperm and angiosperm families. Our strategy combined tissue-specific metabolite analyses with development and mining of deep transcriptome resources using customized reference sequence databases for relevant target gene modules (diTPSs and P450s). Informative phylogenies, from which general patterns of evolution of gene functions and lineage-specific blooms could be inferred, proved useful to guide functional annotation as shown for a set of diTPSs using both in vitro and in vivo analyses of functions. We also showed the potential for developing combinatorial platforms for heterologous production of natural and nonnatural diterpenes, which we had previously proposed as a concept for synthetic biology of plant bioproducts (Facchini et al., 2012). The results of this study provide a resource for much future work in the species of our focus, which are broadly of interest because of the various bioproduct applications of their diterpene metabolites. The diTPS and P450 query databases used in this work, the approaches of identifying candidate genes from functionally informative gene family phylogenies, and the in vivo and in vitro expression systems can be applied to other plant species that produce interesting diterpenes, where scale of investigation can be focused on one particular species of interest or can be a broad-scale search for new enzymes across the plant kingdom. The general model of the present work can also be applied to other classes of terpenoids.

We showed that confirming presence of target compounds by metabolite profiling of different tissues provided confidence in the choice of tissues selected for transcriptome analysis. Since metabolite transport and accumulation in tissues other than their biosynthetic origin cannot be ignored, assessing transcriptomes for core precursor pathways provided

additional confidence in the assembled transcriptome resources, as shown here with the complete coverage of all steps of the MVA and MEP pathway genes in each transcriptome paralleled with large numbers of relevant diTPS and P450 candidates. We found that combining Illumina and 454 assemblies was advantageous as it expanded the recovery of FL assemblies with different read-depth distribution from both sequencing platforms; however, read-length and -depth advantages of different high-throughput sequencing are rapidly changing and will provide continuously improving coverage for targeted gene discovery. The development and use of nonredundant diTPS and P450 databases, which encompass the functional space of these gene families was most important as it allowed efficient query of large transcriptome datasets. In all 10 plant species (except for *R. officinalis*, showing a lower assembly quality) multigene diTPS and P450 families were identified, supporting the identification of patterns of lineage-specific expansion and diversification of the diTPS- and P450-ome in species of different angiosperm and gymnosperm families. Blooms of diTPSs and P450s in species producing diverse diterpene metabolites are hot spots for functional gene identification. Finally, reliable functional annotation is a key challenge in the discovery of diTPS and P450 genes, given the functional plasticity and high sequence similarity within these large gene families. For example, with diTPSs, change of function as a result of small changes in the active site composition has been illustrated (e.g. Wilderman and Peters, 2007; Xu et al., 2007b; Keeling et al., 2008; Morrone et al., 2008; Leonard et al., 2010; Zhou and Peters, 2011; Criswell et al., 2012; Gao et al., 2012; Zerbe et al., 2012b). However, emerging patterns of functional evolution, such as changes in domain architecture and presence or absence of active site signature motifs, can now guide functional annotation and characterization of new diTPSs.

MATERIALS AND METHODS

Plant Collection

Abies balsamea and *Pseudolarix amabilis* were purchased from Arbutus Grove Nursery and Forestfarm. Both species were obtained as 2-year-old trees and maintained in University of British Columbia greenhouses. *Tripterygium wilfordii* was provided by the University of British Columbia Botanical Garden. Seeds of *Coleus forskohlii*, *Rosmarinus officinalis*, *Isodon rubescens*, *Marubium vulgare*, *Euphorbia peplus*, *Jatropha gossypifolia*, and *Grindelia robusta* were obtained from B&T World Seeds and Horizon Herbs, and cultivated under long-day conditions (16-h light; day/night temperature 22°C/25°C) for 2 to 4 weeks and further maintained under greenhouse conditions.

Diterpene Profiling

For the isolation of diterpene metabolites, tissues were ground in liquid nitrogen and extracted at room temperature using suitable solvents. Extracts were filtered through cheesecloth, freed from surplus water by adding anhydrous Na₂SO₄, and reduced under nitrogen stream to approximately 1 mL. Further chromatographic purification on silica, alumina, or charcoal material was applied as required. Samples were filtered through 0.22-μm GHP membrane filters (www.pall.com) and analyzed by LC-MS or GC-MS with

metabolite identification by comparison with authentic standards and reference mass spectra from the Wiley Registry Mass Spectral Libraries. Detailed extraction and analytical procedures are given in Supplemental Materials and Methods S1.

RNA Isolation, cDNA Library Construction, and Transcriptome Sequencing

RNA was isolated as previously described (Kolosova et al., 2004) using 50 to 150 mg of tissue. High RNA integrity was verified using Bioanalyzer 2100 RNA Nano-chip assays (www.chem.agilent.com). Construction of non-normalized cDNA libraries and transcriptome sequencing was performed at the McGill University and Génome Québec Innovation Centre. cDNA libraries for 454 sequencing were constructed from 200 ng of fragmented mRNA using the cDNA Rapid Library Preparation kit, GS FLX Titanium series (www.roche.com). After yield and fragment size validation on a 6,000 Nano-chip (Agilent), 200 ng of prepared libraries were subject to a one-half-plate reaction of 454 pyrosequencing with the Roche GS FLX Titanium platform. For Illumina sequencing, cDNA libraries were prepared from 10 μg of total RNA using the mRNA Seq Sample Preparation Kit (www.illumina.com), with normalization of mRNA amounts to 100 ng prior to library construction. Yields and correct fragment sizes were assessed using a high sensitivity DNA chip (Agilent). Sequencing was performed from 7 pmol of each library using 108 bp (GAIIx) or 100 bp (HiSeq) pair-ended runs.

De Novo Transcriptome Assembly

454 reads were cleaned to remove adapter sequences and 15 bp at front ends of sequences using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Flowgram files were clipped with tools provided by Roche. The resulting high-quality reads were assembled with the “cdna” switch paired with 45-bp minimum read length using the Roche Newbler 2.6 genome assembler. Illumina assemblies were conducted with the Trinity de novo assembler (Grabherr et al., 2011) after cleaning of sequence reads with a custom perl script, allowing trimming of Illumina reads in fastq format and further removal of 15 bp at front ends of sequences. Mapping of raw Illumina reads to respective assemblies was achieved using Burrows-Wheeler Aligner (Li and Durbin, 2009) with default parameters.

Terpenoid Pathway Gene Discovery

Functional annotation of terpenoid backbone biosynthetic genes was conducted using a BLASTx approach against a manually curated version of the terpenoid backbone biosynthesis pathway (ec00900) from the KEGG database with exclusion of duplicates. Assemblies were screened with an *E*-value cutoff of 1e⁻²⁰. For identifying diTPS and P450 candidates, custom databases (diTPS, Supplemental Data S11; P450, Supplemental Data S12) were created based on publicly available protein sequences that represent minimal nonredundant sequence sets, resembling the functional range of TPS and P450 genes. Gene mining of significant candidates from generated assemblies was performed using tBLASTx with an *E*-value threshold of 1e⁻⁵⁰ and a minimum read length of 150 amino acids.

Phylogenetic Analysis

Protein sequence alignments were generated using Dialign 2.2.1 (Morgenstern, 2004) and curated with GBLOCKS version 0.19b (Castresana, 2000). Phylogenetic analyses were performed using a maximum likelihood algorithm in the PhyML-aBayes version 3.0.1 beta (Anisimova et al., 2011) with four rate substitution categories, LG substitution model, BIONJ starting tree, and 100 bootstrap repetitions.

cDNA Cloning

Cloning of FL cDNAs was conducted using gene-specific oligonucleotides (Supplemental Table S5) or, where required, through completion of 5'- and 3'-sequences by RACE using the SMARTer RACE cDNA amplification kit (www.clontech.com) and touchdown PCR from cDNA prepared with the SuperScript III First Strand Synthesis kit (www.invitrogen.com) and random hexamer oligonucleotides. For expression in *Escherichia coli*, obtained amplicons were ligated into pJET (Clontech), sequence verified, and subcloned into

the pET28b(+) expression vector (www.emdmillipore.ca). For alternative expression in *Nicotiana benthamiana*, genes were amplified from FL cDNA clones using PfuX7 polymerase (Nørholm, 2010) and cloned into pCambia130035Su as described earlier (Nour-Eldin et al., 2006).

Expression of diTPS in *E. coli* and diTPS in Vitro Assays

Recombinant proteins were expressed in *E. coli* BL21DE3-C41 and Ni²⁺-affinity purified as described elsewhere (Zerbe et al., 2012b). In vitro enzyme assays were conducted as previously reported (Zerbe et al., 2012b) in the form of single or coupled assays using 50 µg of purified protein (50 µg each for coupled assays) and 15 µM of GPP 1 (Sigma), with incubation for 1 h at 30°C and extraction of reaction products with 500 µL pentane prior to GC-MS analysis. The detection of diphosphate intermediates required treatment with 10 units of calf intestinal phosphatase (Invitrogen) for 16 h at 37°C prior to GC-MS analysis. Detailed procedures are given in Supplemental Materials and Methods S2.

Expression of diTPS in *N. benthamiana* and diTPS in Vivo Assays

For expression in *N. benthamiana*, diTPS candidates were cloned into the pCambia130035Su vector (Nour-Eldin et al., 2006), and transformed into *Agrobacterium tumefaciens* strain GV3850 by electroporation. The resulting strains were grown at 28°C in Luria-Bertani media supplemented with 50 mg L⁻¹ kanamycin, 50 mg L⁻¹ ampicillin and 34 mg L⁻¹ rifampicin. After harvest, cells were resuspended to a final optical density at 600 nm of 2 in 10 mM MES buffer with 10 mM MgCl₂ and 100 µM acetosyringone. Following 60-min shaking incubation at 50 rpm and room temperature, strains were mixed with 0.25 volume of the P19 suppressor strain (Voinnet et al., 2003) for single-construct transformations, or using equal volumes for combinatorial expression of two diTPSs with 0.25 volume of P19 before infiltration into the underside of 5-week-old *N. benthamiana* leaves. Plants were incubated 4 d prior to metabolite extraction with 1 mL hexane and GC-MS analysis as described in Supplemental Materials and Methods S2.

Assemblies produced from the 454 and Illumina sequence reads are available as Supplemental Data: *A. balsamea*: Supplemental Data S1, *P. amabilis*: Supplemental Data S2, *G. robusta*: Supplemental Data S3, *C. forskohlii*: Supplemental Data S4, *R. officinalis*: Supplemental Data S5, *M. vulgare*: Supplemental Data S6, *J. gossypifolia*: Supplemental Data S7, *T. wilfordii*: Supplemental Data S8, *E. peplus*: Supplemental Data S9, *I. rubescens*: Supplemental Data S10.

The 454- and Illumina-derived nucleotide sequences reported in this paper have been submitted to the short read archive (SRA) at NCBI with the following accession numbers: *A. balsamea* (SRX039631 and SRX202899), *P. amabilis* (SRX079089 and SRX096935), *G. robusta* (SRR835835318 and SUB167813), *C. forskohlii* (SRX079017 and SRX235906), *R. officinalis* (SRX079014), *M. vulgare* (SRX079090 and SRX096933), *J. gossypifolia* (SRX130845), *T. wilfordii* (SRX039632 and SRX202900), *E. peplus* (SRX039633 and SRX096931), and *I. rubescens* (SRX079018 and SRX23528). Nucleotide sequences of characterized enzymes have been submitted to the GenBankTM/EBI Data Bank with accession numbers CFTPS14, KC7022394; EpTPS1, KC7022395; EpTPS7, KC7022396; EpTPS3, KC7022397; PxTPS4, KC7022398; GrTPS6, KC7022399; GrTPS1, KC7022400; and GrTPS5, KC7022401.

Supplemental Data

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

Supplemental Figure S1. Average contig length distribution of assemblies generated from 454- and Illumina-derived transcriptomes.

Supplemental Figure S2. Discovery of diTPSs and P450s in the transcriptome assemblies.

Supplemental Table S1. Summary of Roche 454 and Illumina GAI and HiSeq transcriptome sequencing results.

Supplemental Table S2. Comparison of 454 and Illumina assemblies to the Arabidopsis (*Arabidopsis thaliana*) CEGMA data set.

Supplemental Table S3. Abbreviations and accession numbers of diTPSs used for phylogenetic analyses in Figure 4.

Supplemental Table S4. Abbreviations and accession numbers of P450s used for phylogenetic analyses in Figure 5.

Supplemental Table S5. Oligonucleotides used in PCR procedures described in Figures 6 and 7.

Supplemental Data S1. De novo 454 and Illumina assemblies of *A. balsamea* bark tissue.

Supplemental Data S2. De novo 454 and Illumina assemblies of *P. amabilis* root tissue.

Supplemental Data S3. De novo 454 and Illumina assemblies of *G. robusta* flower tissue.

Supplemental Data S4. De novo 454 and Illumina assemblies of *C. forskohlii* root tissue.

Supplemental Data S5. De novo 454 and Illumina assemblies of *R. officinalis* leaf tissue.

Supplemental Data S6. De novo 454 and Illumina assemblies of *M. vulgare* leaf tissue.

Supplemental Data S7. De novo 454 and Illumina assemblies of *J. gossypifolia* root tissue.

Supplemental Data S8. De novo 454 and Illumina assemblies of *T. wilfordii* leaf tissue.

Supplemental Data S9. De novo 454 and Illumina assemblies of *E. peplus* leaf tissue.

Supplemental Data S10. De novo 454 and Illumina assemblies of *I. rubescens* leaf tissue.

Supplemental Data S11. TPS reference database.

Supplemental Data S12. P450 reference database.

Supplemental Materials and Methods S1. Detailed description of extraction and metabolite analysis procedures used for the identification of the respective diterpenes targeted in this study.

Supplemental Materials and Methods S2. Detailed description of GC-MS analyses used for in vitro and in vivo functional characterization of diTPSs as illustrated in Figures 6 and 7.

ACKNOWLEDGMENTS

We thank Ms. Karen Reid (University of British Columbia) for outstanding laboratory and project management; Dr. David Nelson (University of Tennessee) for assistance with the development of the P450 reference database, P450 naming, and helpful discussion; and Ms. Lea Gram Hansen (Copenhagen University) for technical assistance. Transcriptome sequencing was performed at the McGill University and Génome Québec Innovation Centre. This work was part of the PhytoMetaSyn Project.

Received March 21, 2013; accepted April 21, 2013; published April 23, 2013.

LITERATURE CITED

- Ajikumar PK, Xiao WH, Tyo KE, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science* **330**: 70–74
- Alasbahi RH, Melzig MF (2010a) *Plectranthus barbatus*: a review of phytochemistry, ethnobotanical uses and pharmacology - Part 1. *Planta Med* **76**: 653–661
- Alasbahi RH, Melzig MF (2010b) *Plectranthus barbatus*: a review of phytochemistry, ethnobotanical uses and pharmacology - Part 2. *Planta Med* **76**: 753–765
- Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol* **60**: 685–699
- Baldwin IT (2010) Plant volatiles. *Curr Biol* **20**: R392–R397

- Bohlmann J, Meyer-Gauen G, Croteau R** (1998a) Plant terpenoid syntheses: molecular biology and phylogenetic analysis. *Proc Natl Acad Sci USA* **95**: 4126–4133
- Bohlmann J, Crock J, Jetter R, Croteau R** (1998b) cDNA cloning, characterization, and functional expression of wound-inducible (*E*)- α -bisabolene synthase from grand fir (*Abies grandis*). *Proc Natl Acad Sci USA* **95**: 6756–6761
- Bohlmann J, Keeling CI** (2008) Terpenoid biomaterials. *Plant J* **54**: 656–669
- Brinker AM, Ma J, Lipsky PE, Raskin I** (2007) Medicinal chemistry and pharmacology of genus *Tripterygium* (Celastraceae). *Phytochemistry* **68**: 732–766
- Caniard A, Zerbe P, Legrand S, Cohade A, Valot N, Magnard JL, Bohlmann J, Legendre L** (2012) Discovery and functional characterization of two diterpene synthases for sclareol biosynthesis in *Salvia sclarea* (L.) and their relevance for perfume manufacture. *BMC Plant Biol* **12**: 119
- Castresana J** (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552
- Chen F, Tholl D, Bohlmann J, Pichersky E** (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J* **66**: 212–229
- Chiu P, Leung LT, Ko BCB** (2010) Pseudolaric acids: isolation, bioactivity and synthetic studies. *Nat Prod Rep* **27**: 1066–1083
- Criswell J, Potter K, Shephard F, Beale MH, Peters RJ** (2012) A single residue change leads to a hydroxylated product from the class II diterpene cyclization catalyzed by abietadiene synthase. *Org Lett* **14**: 5828–5831
- Davis EM, Croteau R** (2000) Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. *Top Curr Chem* **209**: 53–95
- De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y** (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci USA* **110**: 2898–2903
- Dudareva N, Cseke L, Blanc VM, Pichersky E** (1996) Evolution of floral scent in *Clarkia*: novel patterns of S-linalool synthase gene expression in the *C. breweri* flower. *Plant Cell* **8**: 1137–1148
- Ennajdaoui H, Vachon G, Giacalone C, Besse I, Sallaud C, Herzog M, Tissier A** (2010) Trichome specific expression of the tobacco (*Nicotiana sylvestris*) membratien-ol synthase genes is controlled by both activating and repressing *cis*-regions. *Plant Mol Biol* **73**: 673–685
- Facchini PJ, Bohlmann J, Covello PS, De Luca V, Mahadevan R, Page JE, Ro DK, Sensen CW, Storms R, Martin VJ** (2012) Synthetic biosystems for the production of high-value plant metabolites. *Trends Biotechnol* **30**: 127–131
- Falara V, Pichersky E, Kanellis AK** (2010) A copal-8-ol diphosphate synthase from the angiosperm *Cistus creticus* subsp. *creticus* is a putative key enzyme for the formation of pharmacologically active, oxygen-containing labdane-type diterpenes. *Plant Physiol* **154**: 301–310
- Feyereisen R** (2011) Arthropod CYPomes illustrate the tempo and mode in P450 evolution. *Biochim Biophys Acta* **1814**: 19–28
- Gao W, Hillwig ML, Huang L, Cui G, Wang X, Kong J, Yang B, Peters RJ** (2009) A functional genomics approach to tanshinone biosynthesis provides stereochemical insights. *Org Lett* **11**: 5170–5173
- Gao Y, Honzatko RB, Peters RJ** (2012) Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat Prod Rep* **29**: 1153–1175
- Goyal SK, Samsher, Goyal RK** (2010) Stevia (*Stevia rebaudiana*) a bio-sweetener: a review. *Int J Food Sci Nutr* **61**: 1–10
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652
- Guerra-Bubb J, Croteau R, Williams RM** (2012) The early stages of taxol biosynthesis: an interim report on the synthesis and identification of early pathway metabolites. *Nat Prod Rep* **29**: 683–696
- Hall DE, Zerbe P, Jancsik S, Quesada AL, Dullat H, Madilao LL, Yuen M, Bohlmann J** (2013) Evolution of conifer diterpene synthases: diterpene resin acid biosynthesis in lodgepole pine and jack pine involves monofunctional and bifunctional diterpene synthases. *Plant Physiol* **161**: 600–616
- Hamberger B, Ohnishi T, Hamberger B, Séguin A, Bohlmann J** (2011) Evolution of diterpene metabolism: Sitka spruce CYP720B4 catalyzes multiple oxidations in resin acid biosynthesis of conifer defense against insects. *Plant Physiol* **157**: 1677–1695
- Hamberger B, Bak S** (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120426
- Harris LJ, Saparno A, Johnston A, Prisc S, Xu M, Allard S, Kathiresan A, Ouellet T, Peters RJ** (2005) The maize An2 gene is induced by *Fusarium* attack and encodes an *ent*-copalyl diphosphate synthase. *Plant Mol Biol* **59**: 881–894
- Hayashi K-I, Kawaide H, Notomi M, Sakigi Y, Matsuo A, Nozaki H** (2006) Identification and functional analysis of bifunctional *ent*-kaurene synthase from the moss *Physcomitrella patens*. *FEBS Lett* **580**: 6175–6181
- Heiling S, Schuman MC, Schoettner M, Mukerjee P, Berger B, Schneider B, Jassbi AR, Baldwin IT** (2010) Jasmonate and ppHsystemin regulate key malonylation steps in the biosynthesis of 17-hydroxygeranylinalool diterpene glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. *Plant Cell* **22**: 273–292
- Herde M, Gärtner K, Köllner TG, Fode B, Boland W, Gershenzon J, Gatz C, Tholl D** (2008) Identification and regulation of TPS04/GES, an *Arabidopsis* geranylinalool synthase catalyzing the first step in the formation of the insect-induced volatile C16-homoterpene TMTT. *Plant Cell* **20**: 1152–1168
- Hezari M, Lewis NG, Croteau R** (1995) Purification and characterization of taxa-4(5),11(12)-diene synthase from Pacific yew (*Taxus brevifolia*) that catalyzes the first committed step of taxol biosynthesis. *Arch Biochem Biophys* **322**: 437–444
- Higashi Y, Saito K** (2013) Network analysis for gene discovery in plant specialized metabolism. *Plant Cell Environ* (in press)
- Hillwig ML, Xu M, Toyomasu T, Tiernan MS, Wei G, Cui G, Huang L, Peters RJ** (2011) Domain loss has independently occurred multiple times in plant terpene synthase evolution. *Plant J* **68**: 1051–1060
- Jassbi AR, Gase K, Hettenhausen C, Schmidt A, Baldwin IT** (2008) Silencing geranylgeranyl diphosphate synthase in *Nicotiana attenuata* dramatically impairs resistance to tobacco hornworm. *Plant Physiol* **146**: 974–986
- Jennewein S, Croteau R** (2001) Taxol: biosynthesis, molecular genetics, and biotechnological applications. *Appl Microbiol Biotechnol* **57**: 13–19
- Jennewein S, Long RM, Williams RM, Croteau R** (2004) Cytochrome p450 taxadiene 5 α -hydroxylase, a mechanistically unusual monooxygenase catalyzing the first oxygenation step of taxol biosynthesis. *Chem Biol* **11**: 379–387
- Jiang M, Stephanopoulos G, Pfeifer BA** (2012) Downstream reactions and engineering in the microbially reconstituted pathway for Taxol. *Appl Microbiol Biotechnol* **94**: 841–849
- Keeling CI, 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
- Keeling CI, Weisshaar S, Lin RPC, Bohlmann J** (2008) Functional plasticity of paralogous diterpene synthases involved in conifer defense. *Proc Natl Acad Sci USA* **105**: 1085–1090
- Keeling CI, Dullat HK, Yuen M, Ralph SG, Jancsik S, Bohlmann J** (2010) Identification and functional characterization of monofunctional *ent*-copalyl diphosphate and *ent*-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. *Plant Physiol* **152**: 1197–1208
- Keeling CI, Weisshaar S, Ralph SG, Jancsik S, Hamberger B, Dullat HK, Bohlmann J** (2011a) Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*Picea* spp.). *BMC Plant Biol* **11**: 11–43
- Keeling CI, Madilao LL, Zerbe P, Dullat HK, Bohlmann J** (2011b) The primary diterpene synthase products of *Picea abies* levopimaradiene/abietadiene synthase (PaLAS) are epimers of a thermally unstable diterpenol. *J Biol Chem* **286**: 21145–21153
- Kirby J, Nishimoto M, Park JG, Withers ST, Nowroozi F, Behrendt D, Rutledge EJ, Fortman JL, Johnson HE, Anderson JV, et al** (2010) Cloning of casbene and neocembrene synthases from *Euphorbiaceae* plants and expression in *Saccharomyces cerevisiae*. *Phytochemistry* **71**: 1466–1473
- Köksal M, Jin Y, Coates RM, Croteau R, Christianson DW** (2011) Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **469**: 116–120
- Kolosova N, Miller B, Ralph S, Ellis BE, Douglas C, Ritland K, Bohlmann J** (2004) Isolation of high-quality RNA from gymnosperm and angiosperm trees. *Biotechniques* **36**: 821–824

- Leonard E, Ajikumar PK, Thayer K, Xiao WH, Mo JD, Tidor B, Stephanopoulos G, Prather KL (2010) Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc Natl Acad Sci USA* **107**: 13654–13659
- Li CY, Wang EQ, Cheng Y, Bao JK (2011) Oridonin: An active diterpenoid targeting cell cycle arrest, apoptotic and autophagic pathways for cancer therapeutics. *Int J Biochem Cell Biol* **43**: 701–704
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760
- Li L, Shukla S, Lee A, Garfield SH, Maloney DJ, Ambudkar SV, Yuspa SH (2010) The skin cancer chemotherapeutic agent ingenol-3-angelate (PEP005) is a substrate for the epidermal multidrug transporter (ABC1) and targets tumor vasculature. *Cancer Res* **70**: 4509–4519
- López-Jiménez A, García-Caballero M, Medina MA, Quesada AR (2011) Anti-angiogenic properties of carnosol and carnosic acid, two major dietary compounds from rosemary. *Eur J Nutr* **52**: 85–95
- Ma Y, Yuan L, Wu B, Li X, Chen S, Lu S (2012) Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. *J Exp Bot* **63**: 2809–2823
- Mafu S, Hillwig ML, Peters RJ (2011) A novel labda-7,13e-dien-15-ol-producing bifunctional diterpene synthase from *Selaginella moellendorffii*. *ChemBioChem* **12**: 1984–1987
- Martin DM, Fäldt J, Bohlmann J (2004) Functional characterization of nine Norway spruce *TPS* genes and evolution of gymnosperm terpene synthases of the *TPS-d* subfamily. *Plant Physiol* **135**: 1908–1927
- Martin DM, Aubourg S, Schouwey MB, Daviet L, Schalk M, Toub O, Lund ST, Bohlmann J (2010) Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLDNA cloning, and enzyme assays. *BMC Plant Biol* **10**: 226
- Mau CJ, West CA (1994) Cloning of casbene synthase cDNA: evidence for conserved structural features among terpenoid cyclases in plants. *Proc Natl Acad Sci USA* **91**: 8497–8501
- McAndrew RP, Peralta-Yahya PP, DeGiovanni A, Pereira JH, Hadi MZ, Keasling JD, Adams PD (2011) Structure of a three-domain sesquiterpene synthase: a prospective target for advanced biofuels production. *Structure* **19**: 1876–1884
- Meyre-Silva C, Cechinel-Filho V (2010) A review of the chemical and pharmacological aspects of the genus *marrubium*. *Curr Pharm Des* **16**: 3503–3518
- Mnonopi N, Levendal RA, Mzilikazi N, Frost CL (2012) Marrubiin, a constituent of *Leonotis leonurus*, alleviates diabetic symptoms. *Phyto-medicine* **19**: 488–493
- Morgenstern B (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res* **32**: W33–36
- Morrone D, Xu M, Fulton DB, Determan MK, Peters RJ (2008) Increasing complexity of a diterpene synthase reaction with a single residue switch. *J Am Chem Soc* **130**: 5400–5401
- Nelson D, Werck-Reichhart D (2011) A P450-centric view of plant evolution. *Plant J* **66**: 194–211
- Norholm MH (2010) A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol* **10**: 21
- Nour-Eldin HH, Hansen BG, Norholm MH, Jensen JK, Halkier BA (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res* **34**: e122
- Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**: 1061–1067
- Peters RJ, Flory JE, Jetter R, Ravn MM, Lee HJ, Coates RM, Croteau RB (2000) Abietadiene synthase from grand fir (*Abies grandis*): characterization and mechanism of action of the “pseudomature” recombinant enzyme. *Biochemistry* **39**: 15592–15602
- Peters RJ (2006) Uncovering the complex metabolic network underlying diterpenoid phytoalexin biosynthesis in rice and other cereal crop plants. *Phytochemistry* **67**: 2307–2317
- Peters RJ (2010) Two rings in them all: the labdane-related diterpenoids. *Nat Prod Rep* **27**: 1521–1530
- Pichersky E, Lewinsohn E (2011) Convergent evolution in plant specialized metabolism. *Annu Rev Plant Biol* **62**: 549–566
- Ralston L, Kwon ST, Schoenbeck M, Ralston J, Schenk DJ, Coates RM, Chappell J (2001) Cloning, heterologous expression, and functional characterization of 5-*epi*-aristolochene-1,3-dihydroxylase from tobacco (*Nicotiana tabacum*). *Arch Biochem Biophys* **393**: 222–235
- Reiling KK, Yoshikuni Y, Martin VJJ, Newman J, Bohlmann J, Keasling JD (2004) Mono and diterpene production in *Escherichia coli*. *Biotechnol Bioeng* **87**: 200–212
- Ro DK, Arimura G, Lau SY, Piers E, Bohlmann J (2005) Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. *Proc Natl Acad Sci USA* **102**: 8060–8065
- Sallaud C, Giacalone C, Töpfer R, Goepfert S, Bakaher N, Rösti S, Tissier A (2012) Characterization of two genes for the biosynthesis of the labdane diterpene Z-abienol in tobacco (*Nicotiana tabacum*) glandular trichomes. *Plant J* **72**: 1–17
- Schalk M, Pastore L, Mirata MA, Khim S, Schouwey M, Deguerry F, Pineda V, Rocci L, Daviet L (2012) Toward a biosynthetic route to sclareol and amber odorants. *J Am Chem Soc* **134**: 18900–18903
- Schmelz EA, Kaplan F, Huffaker A, Dafoe NJ, Vaughan MM, Ni X, Rocca JR, Alborn HT, Teal PE (2011) Identity, regulation, and activity of inducible diterpenoid phytoalexins in maize. *Proc Natl Acad Sci USA* **108**: 5455–5460
- Seki H, Ohyama K, Sawai S, Mizutani M, Ohnishi T, Sudo H, Akashi T, Aoki T, Saito K, Muranaka T (2008) Licorice β -amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. *Proc Natl Acad Sci USA* **105**: 14204–14209
- Seo S, Gomi K, Kaku H, Abe H, Seto H, Nakatsu S, Neya M, Kobayashi M, Nakaho K, Ichinose Y, et al (2012) Identification of natural diterpenes that inhibit bacterial wilt disease in tobacco, tomato and *Arabidopsis*. *Plant Cell Physiol* **53**: 1432–1444
- Stofer Vogel B, Wildung MR, Vogel G, Croteau R (1996) Abietadiene synthase from grand fir (*Abies grandis*): cDNA isolation, characterization, and bacterial expression of a bifunctional diterpene cyclase involved in resin acid biosynthesis. *J Biol Chem* **271**: 23262–23268
- Sun TP, Kamiya Y (1994) The *Arabidopsis* GA1 locus encodes the cyclase *ent*-kaurene synthetase A of gibberellin biosynthesis. *Plant Cell* **6**: 1509–1518
- Swaminathan S, Morrone D, Wang Q, Fulton DB, Peters RJ (2009) CYP76M7 is an *ent*-cassadiene C11 α -hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**: 3315–3325
- Theoduloz C, Rodríguez JA, Pertino M, Schmeda-Hirschmann G (2009) Antiproliferative activity of the diterpenes jatrophone and jatropholone and their derivatives. *Planta Med* **75**: 1520–1522
- Voinnet O, Rivas S, Mestre P, Baulcombe D (2003) An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant J* **33**: 949–956
- Walker K, Croteau R (2001) Taxol biosynthetic genes. *Phytochemistry* **58**: 1–7
- Wang E, Wagner GJ (2003) Elucidation of the functions of genes central to diterpene metabolism in tobacco trichomes using posttranscriptional gene silencing. *Planta* **216**: 686–691
- Wang Q, Hillwig ML, Peters RJ (2011) CYP99A3: functional identification of a diterpene oxidase from the momilactone biosynthetic gene cluster in rice. *Plant J* **65**: 87–95
- Wang Q, Hillwig ML, Okada K, Yamazaki K, Wu Y, Swaminathan S, Yamane H, Peters RJ (2012a) Characterization of CYP76M5-8 indicates metabolic plasticity within a plant biosynthetic gene cluster. *J Biol Chem* **287**: 6159–6168
- Wang Q, Hillwig ML, Wu Y, Peters RJ (2012b) CYP701A8: a rice *ent*-kaurene oxidase paralog diverted to more specialized diterpenoid metabolism. *Plant Physiol* **158**: 1418–1425
- Wilderman PR, Peters RJ (2007) A single residue switch converts abietadiene synthase into a pimaradiene specific cyclase. *J Am Chem Soc* **129**: 15736–15737
- Wildung MR, Croteau R (1996) A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of taxol biosynthesis. *J Biol Chem* **271**: 9201–9204
- Wilson SA, Roberts SC (2012) Recent advances towards development and commercialization of plant cell culture processes for the synthesis of biomolecules. *Plant Biotechnol J* **10**: 249–268
- Wu Y, Hillwig ML, Wang Q, Peters RJ (2011) Parsing a multifunctional biosynthetic gene cluster from rice: Biochemical characterization of CYP71Z6 & 7. *FEBS Lett* **585**: 3446–3451
- Wu Y, Zhou K, Toyomasu T, Sugawara C, Oku M, Abe S, Usui M, Mitsuhashi W, Chono M, Chandler PM, et al (2012) Functional characterization of wheat copalyl diphosphate synthases sheds light on the

- early evolution of labdane-related diterpenoid metabolism in the cereals. *Phytochemistry* **84**: 40–46
- Wüst M, Little DB, Schalk M, Croteau R (2001) Hydroxylation of limonene enantiomers and analogs by recombinant (-)-limonene 3- and 6-hydroxylases from mint (*Mentha*) species: evidence for catalysis within sterically constrained active sites. *Arch Biochem Biophys* **387**: 125–136
- Xu M, Wilderman PR, Morrone D, Xu J, Roy A, Margis-Pinheiro M, Upadhyaya NM, Coates RM, Peters RJ (2007a) Functional characterization of the rice kaurene synthase-like gene family. *Phytochemistry* **68**: 312–326
- Xu M, Wilderman PR, Peters RJ (2007b) Following evolution's lead to a single residue switch for diterpene synthase product outcome. *Proc Natl Acad Sci USA* **104**: 7397–7401
- Yamaguchi S, Sun Tp, Kawaide H, Kamiya Y (1998) The GA2 locus of *Arabidopsis thaliana* encodes *ent*-kaurene synthase of gibberellin biosynthesis. *Plant Physiol* **116**: 1271–1278
- Zabka M, Pavela R, Gabrielova-Slezakova L (2011) Promising antifungal effect of some Euro-Asiatic plants against dangerous pathogenic and toxinogenic fungi. *J Sci Food Agric* **91**: 492–497
- Zerbe P, Chiang A, Yuen M, Hamberger B, Hamberger B, Draper JA, Britton R, Bohlmann J (2012a) Bifunctional *cis*-abienol synthase from *Abies balsamea* discovered by transcriptome sequencing and its implications for diterpenoid fragrance production. *J Biol Chem* **287**: 12121–12131
- Zerbe P, Chiang A, Bohlmann J (2012b) Mutational analysis of white spruce (*Picea glauca*) *ent*-kaurene synthase (PgKS) reveals common and distinct mechanisms of conifer diterpene synthases of general and specialized metabolism. *Phytochemistry* **74**: 30–39
- Zerbe P, Bohlmann J (2013) Bioproducts, biofuels, and perfumes: Conifer terpene synthases and their potential for metabolic engineering. *Recent Adv Phytochem* (in press)
- Zhou K, Peters RJ (2011) Electrostatic effects on (di)terpene synthase product outcome. *Chem Commun (Camb)* **47**: 4074–4080
- Zhou K, Xu M, Tiernan M, Xie Q, Toyomasu T, Sugawara C, Oku M, Usui M, Mitsuhashi W, Chono M, et al (2012a) Functional characterization of wheat *ent*-kaurene(-like) synthases indicates continuing evolution of labdane-related diterpenoid metabolism in the cereals. *Phytochemistry* **84**: 47–55
- Zhou YJ, Gao W, Rong Q, Jin G, Chu H, Liu W, Yang W, Zhu Z, Li G, Zhu G, et al (2012b) Modular pathway engineering of diterpenoid synthases and the mevalonic acid pathway for mitratriene production. *J Am Chem Soc* **134**: 3234–3241
- Zulak KG, Bohlmann J (2010) Terpenoid biosynthesis and specialized vascular cells of conifer defense. *J Integr Plant Biol* **52**: 86–97