Molecular Phylogenetics and Evolution 68 (2013) 432–442

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Diversification of the monoterpene synthase gene family (TPSb) in *Protium*, a highly diverse genus of tropical trees

Felipe Zapata *, Paul V.A. Fine

Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 27 October 2012
Revised 16 April 2013
Accepted 25 April 2013
Available online 9 May 2013

Keywords:
Burseraceae
Codon model
Molecular ecology
Secondary metabolite
Terpenoid

ABSTRACT

Plant monoterpenes are a diverse class of secondary metabolites mediating biotic and abiotic interactions with direct effects on plant fitness. To evaluate the hypothesis that monoterpene diversity is related to functional diversification after gene duplication, we reconstructed the evolutionary history of monoterpene synthases (TPSb) – the genes underlying monoterpene synthesis – in *Protium*, a taxonomically and chemically diverse genus of tropical trees. We isolated multiple copies of TPSb genes from chemically divergent *Protium* species, reconstructed the phylogeny of this gene family, used maximum-likelihood estimation of selection coefficients, and inferred residues evolving under positive selection. We found evidence for one ancient and multiple more recent duplication events giving rise to three, and potentially five, copies of TPSb genes currently present in *Protium*. There was evidence for adaptive evolution in one copy with a positively selected residue likely involved in protein folding and product specificity. All other copies were inferred to be evolving under a combination of stabilizing and/or relaxed selection. Although gene copy number is consistent with the extensive phenotypic diversity in monoterpenes shown in *Protium*, selection analyses suggest that not all copies are undergoing divergent selection consistent with a coevolutionary arms race with enemies, but instead may be under stabilizing and relaxed selection consistent with signaling or physiological stress functionality.

1. Introduction

Monoterpenes are a large group of organic secondary metabolites commonly produced by different taxa, particularly seed plants (Langenheim, 2003). Due to their small molecular weight (10 carbon atoms basic skeleton), monoterpenes are usually emitted as volatiles either as single compounds, components of mixtures, or sequestered as part of resins, which may include other kinds of terpenes, e.g., sesquiterpenes (Trapp and Croteau, 2001a). These emissions play a broad range of ecological roles in nature (Pichersky and Gang, 2000; Gershenzon and Dudareva, 2007), chiefly as direct defenses against herbivores, or indirectly by attracting herbivore enemies (predators and parasitoids) (Arimura et al., 2004; Keeling and Bohlmann, 2006; Mumm and Hilker, 2006; Schnee et al., 2006; Snoeren et al., 2010; Xiao et al., 2012). Besides this role, these emissions can also serve to attract pollinators (Pichersky and Gershenzon, 2002), as communication cues between plants to alert the presence of enemies (Frost et al., 2007; Ton et al., 2006), or to protect plants from abiotic stresses such as the exposure to high temperatures or the oxidative damage due to the accumulation of reactive oxygen (Vickers et al., 2009). Therefore, terpenes mediate interactions that can have direct effects on plant fitness, and thus it is likely that their evolution may be molded by natural selection.

Monoterpene production is catalyzed by monoterpene synthases, a group of enzymes encoded by terpene synthase genes. These genes are part of the highly diverse TPS gene family (Bohlmann et al., 1998; Chen et al., 2011) and play a critical role late in the biosynthetic pathway of terpenes (Davis and Croteau, 2000). Monoterpenes are largely derived from the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway in plastids where monoterpene synthases use geranyl diphosphate (GPP) or neryl diphosphate (NPP)–both derived from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP)–as substrates to generate the enormous diversity of monoterpene carbon skeletons. There is emerging evidence that monoterpenes can also originate from precursors derived in the mevalonate (MVA) pathway in the cytosol in certain taxa (e.g., Hampel et al., 2006). Monoterpene synthases ionize the substrate creating intermediate carbocations that can undergo a series of cyclizations, hydride shifts or other rearrangements before the reaction is terminated. The stochastic nature of these reactions and the fact that an individual carbocation can have multiple fates largely explains why a single enzyme can catalyze the production of multiple monoterpenes (Davis and Croteau, 2000). Thus, monoterpene diversity arises not only because of the large number of different monoterpene synthase

* Corresponding author. Present address: Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912, USA. Fax: +1 401 863 2166.
E-mail address: felipe_zapata@brown.edu (F. Zapata).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.04.024

enzymes currently described (Degenhardt et al., 2009), but also from the biochemical flexibility of each enzyme.

Most angiosperm monoterpene synthases belong to the TPSb subfamily (Bohlmann et al., 1998; Chen et al., 2011). These genes code for proteins around 600–650 amino acids long. They have seven exons of varying length, and they are composed of two distinct structural domains. The N-terminal domain contains a plastid transit peptide and an RRx_8W motif important for monoterpene cyclization (Davis and Croteau, 2000); however, few functional elements have been identified in this domain (e.g., Shishova et al., 2007 for a study on sesquiterpenes). Conversely, the C-terminal domain contains the active site including the DDxxD and NSE/DTE motifs, both involved in ionization of the substrate (Degenhardt et al., 2009). Furthermore, structure–function and mutational studies have shown that mutations at the C-domain can alter the monoterpene products catalyzed by particular TPSb enzymes (El Tamer et al., 2003; Peters and Croteau, 2003; Kampranis et al., 2007). This suggests that functionality of TPSb genes is strongly associated to this domain.

Phylogenetic analyses have shown that TPSb – is closely related to TPSg – (acyclic monoterpenes in angiosperms), TPSa – (sesquiterpene synthases in angiosperms), and TPSd – (mono and sesquiterpenes in gymnosperms) clade genes, and only distantly related to other TPS genes (Bohlmann et al., 1998; Chen et al., 2011). Recent genomic analyses have revealed that different plant taxa harbor multiple closely related copies of TPSb genes – as well as other TPS genes (Aubourg et al., 2002; Martin et al., 2010; Falara et al., 2011). Both of these analyses suggest that the TPS family, as well as all its subfamilies, have a dynamic evolutionary history characterized by both old and recent duplication events followed by sub- and neofunctionalization (e.g., Trapp and Croteau, 2001b; Keeling et al., 2011). These processes can provide the raw material for the diversification of biochemical pathways that may lead to phenotypic diversity (Flagel and Wendel, 2009).

The increasing availability of plant genomes and transcriptomes has provided valuable insights into the evolutionary history, biochemistry, and ecological functions of terpene synthases (e.g., Martin et al., 2010; Chen et al., 2011; Falara et al., 2011; Keeling et al., 2011). However, relatively little work has been carried outside of 'model-plant taxa', particularly on plant groups that produce a large diversity of terpenes and are major components of major ecosystems (but see Keszei et al., 2008, 2010; Keeling et al., 2011). *Protium*, a pantropical genus of approximately 140 species of trees, is an important constituent of the tropical rain forest. It is especially diverse and abundant in the Amazon basin where single sites can harbor up to 35 sympatric species (Rankin-de-Morona et al., 1992; Fine et al., 2005), and it is often in the top five genera in terms of relative basal area of tropical forest plots throughout the basin (Matos and Amaral, 1999; Lima Filho et al., 2001). As a typical member of the family Burseraceae (Weeks et al., 2005), *Protium* is well known for the production of secondary metabolites including different kinds of terpenes (Langenheim, 2003). Almost 100 different mono and sesquiterpenes have been characterized in different species (Siani et al., 2004; Marques et al., 2010; Silva et al., 2009; P.V.A Fine, unpublished). Species within *Protium* produce many mono- and sesquiterpenes, but different species produce different mixtures of these compounds, and some do not produce any monoterpenes at all (Fine et al., 2006; Table 1). In fact, some sections of *Protium* (e.g., Papilloprotium (Daly and Fine, 2011), and Pepeanthos (Daly, 2007)) produce little to no monoterpenes and sesquiterpenes in their leaves and stems, and instead contain a milky white latex. The high degree of variation in monoterpene expression within *Protium* is consistent with the hypothesis that biotic interactions may have selected for divergent chemical defenses and/or signaling compounds in the different species. Therefore, *Protium* represents an excellent opportunity to learn more about natural variation in the genes underlying monoterpene synthesis, and the possible drivers of such variation.

Here, we reconstructed the evolutionary history of monoterpene synthases (TPSb) within *Protium* in the broader context of the TPS family in order to test the hypothesis that TPSb genes have duplicated extensively in *Protium*, and have diversified in function following duplication. In order to test this, we sequenced TPSb genes from representative *Protium* species, generated gene trees, and tested for positive selection on branches and within clades immediately following duplication events. We provide a first glimpse into the evolutionary history of TPSb genes in a chemically and taxonomically diverse group of trees, to gain insights into the molecular bases of biotic interactions, and thereby facilitate the use of these genes in future ecological genetics studies (e.g., Bernhardsson and Ingvarsson, 2012).

2. Materials and methods

2.1. Taxon sampling and isolation of nucleic acids

We generated sequence data for 14 species of *Protium* for this study (Table 1). We chose these species to represent diversity on terpene profiles (P.V.A. Fine, unpublished) as well as to broadly sample the organismal phylogeny of *Protium*, including closely related pairs (e.g., *P. alvarezianum* and *P. subserratum*; Daly and Fine, 2011) and representatives of as many sub-clades as possible (Fine et al., 2005; Daly et al., 2012; P.V.A. Fine, unpublished). We also generated sequence data on one species of *Bursera*, an appropriate close relative outgroup for *Protium* (Weeks et al., 2005). For *Protium*, total genomic DNA (gDNA) was isolated using the DNEasy Plant Mini Kit (Qiagen, Valencia, CA, USA), and for *Bursera*, total RNA was isolated using PureLink Plant RNA Reagent (Invitrogen, Carlsbad, CA, USA). Details of isolation of nucleic acids are provided in Methods A1 (see Supplementary material).

2.2. Primer design

We downloaded complete mRNA sequences from a selection of characterized TPSb genes (Degenhardt et al., 2009; Table A1 of the Supplementary material) for taxa in the Vitales plus Rosidae clade, including *Citrus*, the only other member of the Sapindales clade (Burseraceae also belongs to the Sapindales) with known TPSb gene sequences available in the public NCBI databases. We used gDNA sequences from *Arabidopsis thaliana* to infer likely intron sizes in *Protium* and searched for relatively conserved regions across exons to design degenerate forward and reverse primers. We tried to design primers as close as possible to known functional domains (e.g., DDxxD motif in exon 4) and then extended from these points about 1 kb upstream and downstream (Fig. 1). According to studies on structure–function relationships in TPS genes (e.g., Whittington et al., 2002; El Tamer et al., 2003; Kampranis et al., 2007) this strategy should allow the amplification of the N-domain and functionally relevant sites at the C-domain. The list of primers designed and used in this study is detailed in Table A2 (see Supplementary material).

2.3. PCR, cloning and sequencing

Polymerase chain reaction (PCR) was used to amplify TPSb genes from *Protium* and *Bursera* in two fragments (hereafter, F1 and F2) using the primer pairs B and I, and D and L, respectively (Table A2 of the Supplementary material). Each PCR reaction had a final volume of 25 μl and contained 0.6 pmol of each forward and reverse primers, 2x Green GoTaq Promega reaction buffer, 400 μM dNTPS, 3 mM $MgCl_2$ and 1 unit of GoTaq DNA polymerase

Table 1
Species of *Protium* included in this study with terpene profile, voucher information and GenBank accession number for each fragment and copy sequenced.

Species[a]	Terpene profile	Voucher[b]	Fragment/copy[c]	Accession no.
Protium alstonii (*Pr_altso*)	α-Pinene, β-Pinene, α-Phellandrene, Limonene, γ-Terpinene	Tree 23-IV-6a (AmaLin permanent plot, Perú)	F1C1; F1C2	KC881111; KC881112
	Several sesquiterpenes		F2C4; F2C5	KC881142; KC881143
Protium alvarezianum (*Pr_alvar*)	No monoterpenes	N. Dávila 5954 (Jenaro Herrera, Perú)	F1C2; F1C3	KC881113; KC881114
	Trace sesquiterpenes		F2C2; F2C3; F2C5	KC881144; KC881145; KC881146
Protium apiculatum (*Pr_apicu*)	No monoterpenes	Tree 6-E-14 (AmaLin permanent plot, French Guiana)	F1C3	KC881115
	Trace sesquiterpenes		F2C2	KC881147
Protium calanense (*Pr_calan*)	α-Pinene, β-Pinene, δ-3-Carene.	Tree 12-C-2 (AmaLin permanent plot, French Guiana)	F1C1; F1C3	KC881116; KC881117
	Several sesquiterpenes		F2C3; F2C4; F2C5	KC881148; KC881149; KC881150
Protium decandrum (*Pr_decan*)	α-Pinene, β-Pinene, Myrcene, Limonene	Tree 21-H-9-(AmaLin permanent plot, French Guiana)	F1C1; F1C3	KC881118; KC881119
	Several sesquiterpenes		F2C1; F2C2; F2C4; F2C5	KC881151; KC881152; KC881153; KC881154
Protium giganteum (*Pr_gigan*)	No analyses conducted on leaves. de Freitas et al. (2011) reported terpenes of this species in the essential oils	Tree 18-F-12 (AmaLin permanent plot, French Guiana)	F1C1; F1C3	KC881120; KC881121
			F2C2; F2C4; F2C5	KC881155; KC881156; KC881157
Protium hebetatum (*Pr_hebet*)	α-Pinene, β-Pinene, Limonene	Tree 16-IV-A (AmaLin permanent plot, Perú)	F1C3	KC881122
	Several sesquiterpenes		F2C3; F2C5	KC881158; KC881159
Protium heptaphyllum (*Pr_hepta*)	α-Pinene, β-Pinene, Myrcene, α-Phellandrene, Limonene, γ-Terpinene	PF1421 (Nourages, French Guiana)	F1C2; F1C3	KC881123; KC881124
	Several sesquiterpenes		F2C3; F2C4	KC881160; KC881161
Protium opacum (*Pr_opacu*)	No monoterpenes	PF1418 (Nourages, French Guiana)	F1C1; F1C3	KC881125; KC881126
	Some sesquiterpenes		F2C4; F2C5	KC881162; KC881163
Protium opacum 2 (*Pr_opacu2*)	No monoterpenes	Tree 37-C-12 (AMaLin permanent plot, French Guiana)	F1C1; F1C3	KC881127; KC881128
	Some sesquiterpenes		F2C2; F2C3; F2C4	KC881164; KC881165; KC881166
Protium pallidum (*Pr_palli*)	α-Pinene, Myrcene, α-Phelladrene, δ-3-Carene, Limonene	Tree 12-B-20 (AmaLin permanent plot, French Guiana)	F1C1; F1C2; F1C3	KC881129; KC881130; KC881131
	Several sesquiterpenes		F2C4; F2C5	KC881167; KC881168
Protium pilosum (*Pr_pilos*)	No analyses conducted	PF1452 (Regina, French Guiana)	F1C1; F1C2; F1C3	KC881132; KC881133; KC881134
			F2C2; F2C3; F2C4; F2C5	KC881169; KC881170; KC881171; KC881172
Protium plagiocarpium (*Pr_plagi*)	No analyses conducted	Tree 4-A-27 (AmaLin permanent plot, French Guiana)	F1C1; F1C3	KC881135; KC881136
			F2C2; F2C3; F2C5	KC881173; KC881174; KC881175
Protium subserratum (*Pr_subse*)	No monoterpenes	N. Dávila 5989 (Jenaro Herrera, Perú)	F1C3	KC881137
	Trace sesquiterpenes		F2C5	KC881176
Tetragastris panamensis (*Te_panam*)[d]	No monoterpenes	Tree 4-G-13 (AmaLin permanent plot, French Guiana)	F1C1; F1C2; F1C3	KC881138; KC881139; KC881140
	Some sesquiterpenes		F2C5	KC881177
Busera hindsiana (*Bu_hinds*)	No analyses conducted. Other *Bursera* species are famous for monoterpenes (Evans et al., 2000)	UC920383	AZ1; AZ2	KC881141
				KC881178

[a] Species full name and abbreviation used in figures.
[b] All vouchers are deposited at UC. Trees from permanent plots also have a voucher specimen at UC, but do not have a collection number.
[c] For each fragment (F1 or F2) different copies (C1–C5) were sequenced.
[d] The genus *Tetragastris* is nested within *Protium* and will be transferred (P.V.A. Fine, unpublished data).

(Promega, Madison, WI, USA). PCR thermocycling conditions for both fragments consisted of an initial denaturation step of 30 s at 95 °C followed by 34 cycles of 95 °C for 30 s, 53 °C for 1 min, and 72 °C for 1:30 m, and a final extension step of 72 °C for 7 min. Single band PCR products were visualized on 1% TBE agarose gels and cloned using the pGEM-T vector system (Promega, Madison, WI, USA) following the manufacturer's instructions. Blue–white colony screening was used to pick between 8 and 20 positively

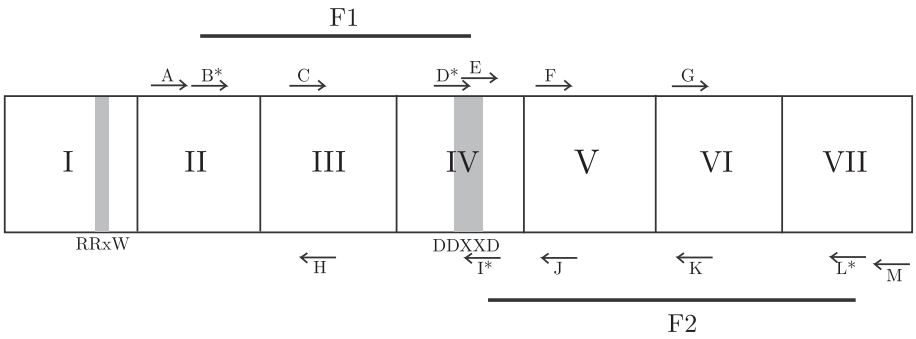

Fig. 1. Schematic representation of a TPSb gene, illustrating gene structure and approximate location of primers designed and used (indicated with *) in this study. Boxes with roman numerals represent exons. Motifs characterizing TPSb genes in exons I and IV are shaded. Thick dark line represent fragments 1 (F1) and 2 (F2) sequenced in this study. Exon sizes are not drawn proportional to size. For variation in exon sizes of TPSb genes see Aubourg et al. (2002).

transformed colonies per PCR product, which were further verified by PCR using universal M13F/R primers. Plasmids were cleaned following the FastPlasmid Mini Kit (Fisher Scientific, Gaithersburg, MD, USA) protocol before sequencing all positive colonies. Sequencing reactions had a final volume of 10 μl and we used the standard Big Dye 3.1 (Applied Biosystems, Foster City, CA, USA) terminator cycle protocol with universal primers SP6-T7. Cloning reactions on *B. hindsiana* failed, thus we used the PCR amplification primer pair for direct sequencing; no polymorphisms were detected in these sequences. Sequencing products were cleaned using an ethanol precipitation step, and analyzed on an ABI 3730 automated DNA Analyzer (Applied Biosystems, Foster City, CA, USA). Geneious Pro 5.5.6 (Drummond et al., 2011) was used to analyze and edit ABI chromatograms, and to assemble double-stranded consensus sequences (>80% overlap) using only high quality, unambiguous reads. All clones per accession were aligned with MUSCLE (Edgar, 2004) with a maximum of 10 iterations and default parameters. Only identical sequences that occurred more than once within the pool of clones per accession were used in subsequent analyses.

2.4. Homology, exon–intron boundaries and multiple sequence alignment

To infer whether we had successfully amplified homologous genes to TPSb, we relied on estimates of statistically significant excess sequence similarity (Pearson and Wood, 2007) using the FASTA package (Pearson and Lipman, 1998) available online (http://fasta.bioch.virginia.edu/fasta_www2). Initially, we queried all gDNA (*Protium*) and cDNA (*Bursera*) sequences against the nucleotide database using the default parameters for sequence comparison (scoring matrix, gap opening/extension). Once introns were excised from gDNA sequences (see below), we queried all translated amino acid sequences against the comprehensive SwissProt and NCBI RefSeq protein databases using the BLOSUM80, BLOSUM62, and BLOSUM50 scoring matrices.

To infer exon–intron boundaries, we first inferred open reading frames (ORFs) in *Bursera* cDNA F1 and F2 by translating nucleotides into amino acids in SeaView 4.3.3 (Guoy et al., 2010) in all frames. We used the ORFs that showed no premature stop codons and separately aligned each fragment with translated *Citrus* cDNA sequences. These alignments were used as reference to manually align each *Protium* gDNA sequence, and we searched for GT(X_n)AG flanking sequences that typically characterize the beginning and end of introns, while checking for ORFs in *Protium* that matched the *Bursera* – *Citrus* alignments. Once we confirmed exon–intron boundaries, we excised introns and only used coding sequences for the rest of this study. Since the F1 and F2 fragments only overlapped by 24 bp, we considered these as two separate data sets.

We downloaded complete mRNA sequences from characterized TPSb genes across the angiosperms (Degenhardt et al., 2009; Table A1 of the Supplementary material), including all gene copies derived from genome-wide analyses in *Arabidopsis* (Aubourg et al., 2002), *Vitis* (Martin et al., 2010), and *Solanum* (Falara et al., 2011). Given that for F1 we did not amplify the conserved RRx_8W motif and the transit peptide region at the N-terminal domain typical in TPSb, we downloaded mRNA from characterized TPSg and TPSe/f genes known to catalyze monoterpene production in other angiosperms (Dudareva et al., 2003; Degenhardt et al., 2009; Martin et al., 2010; Chen et al., 2011) to assess the phylogenetic relationship of F1 in *Protium* to monoterpene synthases of these other TPS subfamilies. In addition, we downloaded sequences from characterized TPSa (angiosperms sesquiterpene synthases) and TPSd (gymnosperms terpene synthases) to serve as outgroups to root our phylogenetic trees. All nucleotide sequences were translated into amino acids and aligned using MAFFT 6.864 (Katoh et al., 2009) with the E-INS-i algorithm, which we ran iteratively through SeaView 4.3.3 (Guoy et al., 2010) to improve alignment quality on several ambiguously aligned regions. After no further improvement was achieved, the resulting protein alignments were translated back to nucleotides and these alignments were inspected by eye to ensure alignment quality before final analyses. The final alignment matrices are deposited in Dryad (http://dx.doi.org/10.5061/dryad.62p02).

2.5. Phylogenetic analyses

To evaluate whether data sets should be partitioned by codon sites or by any combination of codon sites before phylogenetic inference, we used the Bayesian Information Criterion (BIC) (Sullivan and Joyce, 2005) to select the appropriate partitioning scheme and the best-fit nucleotide substitution model for such partitioning using PartitionFinder (Lanfear et al., 2012). For these analyses, we unlinked branch length estimates for each of the substitution models implemented in Mr. Bayes v3 (Ronquist and Huelsenbeck, 2003) in each partition and used the greedy algorithm option. Results suggested that both data sets (F1 and F2) should be treated as a single partition evolving under a GTR + I + Γ model of nucleotide substitution.

We analyzed the sequences as nucleotides using a Bayesian approach as implemented in Mr. Bayes v3.1.2 (Ronquist and Huelsenbeck, 2003), which we ran on the freely available web-based platform Bioportal (http://www.bioportal.uio.no). A summary of the parameters for the likelihood and prior models for both matrices, as well as the settings of the MC^3 (Metropolis-coupled Markov Chain Monte Carlo) sampling are given in Table 2. Details of

Bayesian phylogenetic analyses are provided in Methods A1 (see Supplementary material). A consensus tree was used to summarize the trees sampled during the MC^3. We also analyzed the sequences as codons using a maximum likelihood (ML) optimization framework as implemented in Garli (Zwickl, 2006), and the non-parametric bootstrap (Felsenstein, 1985) to assess branch support with 200 replicates per dataset. For these analyses, we used the codon model of evolution of Goldman and Yang (1994), modeling the substitution process with a two rate parameter, codon frequencies estimated from nucleotide frequencies observed in the data at each codon position (F3x4 model), and variability in the relative rate of nonsynonymous to synonymous substitutions modeled using a discrete distribution with three categories. Details of ML phylogenetic analyses are provided in Methods A1 (see Supplementary material).

2.6. Tests of selection

Structure–function and mutational analyses have shown that the C-domain contains the active site and that most site mutations in this domain can change the terpene profile catalyzed by TPSb genes (El Tamer et al., 2003; Peters and Croteau, 2003; Kampranis et al., 2007), therefore we restricted tests of selection to this domain (i.e., to F2). To quantify selective pressure, we used the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$), with $\omega < 1$, $\omega = 1$, and $\omega > 1$ indicating purifying (or negative) selection, neutral evolution, and diversifying (or positive) selection, respectively (Yang and Bielawski, 2000). For these analyses, we used the codon model specified above for ML phylogenetic reconstruction, and we accounted for variability of selection pressures among codons (also called site models) (Yang et al., 2000). We restricted these analyses only to within each of the *Protium* orthologous copies of F2 with more than one sequence. We included Pr_alvar_F2C5 and Pr_subse_F2C5 in the analysis of C5 because these species grouped within this copy albeit with low support (<0.5 posterior probability, <70% bootstrap; not shown). First, we created alignments for each copy using only *Protium* sequences and excluding other divergent taxa that can reduce the power to detect positive selection with site models (Anisimova et al., 2001). For each of these alignments we estimated unrooted trees using the codon model specified above and a single ω in Garli 2.0 (Zwickl, 2006), which we ran twice with two search replicates each. Each analysis for each alignment converged to the same tree and likelihood score, thus we used these trees and associated branch lengths (in substitutions per codon) as starting trees in maximum likelihood iterations to fit different site models in PAML v.4.4 (Yang, 2007). To test for signals of positive selection, we considered the following models: M1a (nearly neutral), M2a (positive selection), M7 (β) and M8 (β and ω) (Wong et al., 2004). M1a specifies two classes of sites, sites with $0 < \omega < 1$ and neutral sites with $\omega = 1$, whereas M2 adds a third class allowing $\omega > 1$. M7 assumes a β distribution for $0 \leqslant \omega \leqslant 1$, whereas M8 adds an extra category, with a proportion of sites with $\omega > 1$; for M7 and M8, we specified 10 discrete classes of sites. Given that M2 and M8 can produce different estimates of ω depending on the starting values (Wong et al., 2004), we reran these models three times using random starting values for ω and κ (the ratio of transition to transversion rates) drawn from uniform distributions $U(0,5)$ and $U(0,10)$ for ω and κ, respectively. Only the result with the highest log-likelihood values are presented. We used two likelihood ratio tests (LRTs) to test for sites evolving by positive selection comparing (i) M1a (null hypothesis) against M2a, and (ii) M7 (null hypothesis) against M8. Whenever an LRT was significant, we applied the Bayes Empirical Bayes (BEB) (Yang et al., 2005) approach to predict which sites were under positive selection. A site was predicted as positively selected if it belonged to the positive selection category ($\omega > 1$) with posterior probability > 0.95.

To evaluate for signals of positive selection affecting sites along branches (i.e., foreground branches) representing each paralogous TPSb copy in *Protium*, we also optimized two codon models (Zhang et al., 2005). Model A assumes four classes of sites, two containing sites evolving constantly over time under purifying selection or neutrally, and two classes that allow selective pressure at a site to change over time or evolve under positive selection ($\omega > 1$) on foreground branches. We contrasted this model against a null model A′ that does not allow evolution under positive selection ($\omega = 1$). Significance was evaluated with a LRT. For computational efficiency, we restricted this analysis to the sequences in the TPSb clade excluding sequences from all other TPS subfamilies. As for the site models analyses (see above), we estimated an unrooted tree using the codon model specified above and a single ω in Garli 2.0 (Zwickl, 2006), which we ran twice with two search replicates each. Each analysis converged to the same tree (and the same tree as in the ML analysis) and likelihood score, thus we used this tree and associated branch lengths (in substitutions per codon) as starting trees in maximum likelihood iterations to fit branch-sites models in PAML v.4.4 (Yang, 2007). Since branch sites models may have problems of convergence, we reran model A three times using random starting values for ω and κ (the ratio of transition to transversion rates) drawn from uniform distributions $U(0,5)$ and $U(0,10)$ for ω and κ, respectively.

Table 2
Parameter settings for Bayesian phylogenetic analyses for both matrices.

Parameter	Setting
Substitution model	GTR + I + Γ
State frequencies prior	Dirichlet (1,1,1,1)
Substitution matrix prior	Dirichlet (1,1,1,1,1,1)
Alpha shape prior	Exponential (1)
Invariable sites prior	Uniform (0,1)
Branch length prior	Unconstrained: Exponential (10)
Topology prior	Uniform
Number of generations	15,000,000
Sampling frequency	Every 3000
Heating parameter	0.20
Burn-in	20%

3. Results

3.1. Isolation and identification of TPSb genes

PCR amplification using degenerate primers generated a product of approximately 900 bp for each fragment in *Protium*, and 500 bp for each fragment in *Bursera*. Cloning and sequencing of multiple positive colonies (between 8 and 20 for each accession) in *Protium* revealed up to three (F1) and four (F2) considerably different types of sequences per PCR product. For all types of sequences that we recovered in all species, similarity searches using either gDNA or cDNA (after introns were excised; see methods) revealed statistically significant excess similarity to characterized TPSb genes in other angiosperms ($e \ll 10^{-20}$) and to other TPS genes in angiosperms and gymnosperms ($e \ll 10^{-5}$). As expected, when we used scoring matrices with deeper horizons (e.g., BLOSUM50), we were able to recover statistically significant hits to more divergent TPS homologs. These analyses confirmed that we effectively isolated TPSb genes in *Protium* and *Bursera*, and the existence of paralogous copies of TPSb within *Protium*. These copies differed mainly in intron length due to the presence of repeats or indels; differences in exons were largely due to nucleotide substitutions.

In agreement with the highly conserved intron–exon structure of TPSb genes (Trapp and Croteau, 2001b), and given the priming

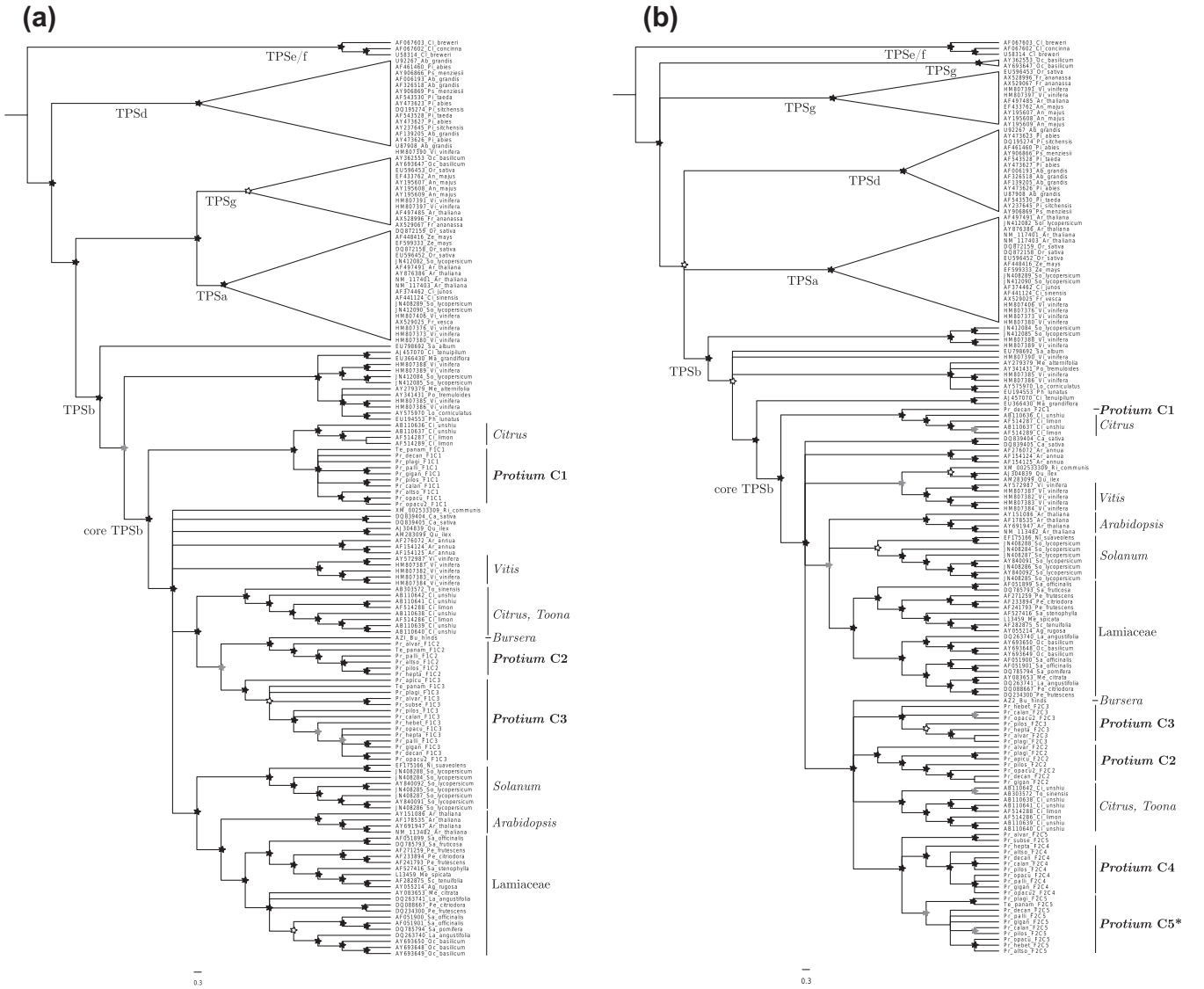

Fig. 2. Angisoperm-wide consensus trees (70%) from Bayesian phylogenetic analyses of terpene synthase genes (TPS) using data as nucleotides for (a) Fragment 1 and (b) Fragment 2. TPS subfamilies indicated under branches. Stars at each node represent posterior probabilities (pp): $1 < pp \leqslant 0.9$ (black); $0.9 < pp \leqslant 0.8$ (gray); $0.8 < pp \leqslant 0.7$ (white); $pp < 0.7$ (no symbol). Clades discussed in the text are indicated with vertical lines. Gene copy with signal of diversifying selection indicated with *.

sites of the primers we designed (Fig. 1), F1 included part of exons two and four, the full length of exon three, and the two intervening intronic regions. Likewise, F2 included part of exons four and seven, the full length of exons five and six, and the three intervening introns. Since F1 and F2 overlapped only around the highly conserved DDxxD domain (i.e., the end of F1 and the start of F2), we decided not to concatenate both fragments even for the same species, because we did not know which paralogous copies in each fragment should be joined.

The structure of F1 and F2 also coincided with the domains and motifs already characterized for TPS genes. F1 largely corresponded to the N-terminal domain of characterized TPS genes (Bateman et al., 2004), except we did not amplify the first exon, and thus we did not sequence the highly conserved RRx_8W motif and the transit peptide region typical in TPSb genes. Nevertheless, multiple sequence alignment revealed consistent positional homology throughout most of the length of F1 with respect to TPSb genes from other angiosperms. F2 corresponded to most of the C-terminal domain of previously characterized TPS genes (Bateman et al., 2004). This domain contained two highly conserved motifs that characterize these genes. All the sequences that we generated contained the DDxxD motif in exon four with no modifications, and contained the NSE/DTE motif in exons six and seven in a modified version (L,Y)(T,Q,S,A)(N,D)D(L,M)xTxxxE.

3.2. Phylogenetic analyses

At a broad scale, phylogenetic analyses of F1 and F2 resulted in congruent topologies using either nucleotide or codon sequence data (Figs. 2 and 3). Codon-based maximum likelihood analyses resulted in poorly supported relationships at deeper nodes, thus these trees were largely congruent with the nucleotide-based Bayesian topologies. The topologies were also consistent with previous phylogenetic studies in the TPS family (e.g., Bohlmann et al., 1998; Chen et al., 2011). All subfamilies that we included in this study were monophyletic with the exception of TPSg where some members were not always part of this clade. The TPSe/f subfamily was sister to the TPSa, TPSd, TPSg and TPSb subfamilies with strong support; however, the relationships among these four subfamilies were not fully congruent between fragments, although support for these nodes in F2 was not high (Figs. 2 and 3).

Within the TPSb subfamily, genes underlying the synthesis of isoprenes and acyclic monoterpenes such as geraniol and (E)-β-ocimene formed a grade sister to the "core TPSb" (Figs. 2 and 3a; Table A1 of the Supplementary material), consistent with previous studies (Sharkey et al., 2005). All the sequences that we generated for *Protium* and *Bursera* were members of the core TPSb clade. Three and five paralogous copies of TPSb genes were recovered in *Protium* using F1 and F2, respectively, suggesting that there may be potentially five copies of TPSb within the *Protium* genome. Unfortunately, it was not possible to recover each and every copy for all species for each fragment. For example, in F1 we recovered copy one (hereafter, C1) in 10 species, whereas in F2 we recovered C1 only for one species (Figs. 2 and 3). Similar cases characterized the other paralogous copies either between or within fragments (i.e., not all paralogous copies of the same fragment were recovered for the same species). C1 was sister to a clade of genes underlying synthesis of the monoterpene (+)-(4R)-limonene in the genus *Citrus*, another member of the order Sapindales (Figs. 2 and 3; Lücker et al., 2002; Shimada et al., 2004, 2005). This clade was sister to the rest of the core TPSb, which was poorly resolved for both fragments. The other paralogous copies (hereafter, C2, C3, C4, and C5) within *Protium* were closely related in all tree topologies and formed a clade with genes underlying synthesis of mostly cyclic monoterpenes from other members of the order Sapindales, including *Citrus* (Rutaceae), *Toona* (Meliaceae), and *Bursera* (Burseraceae). The topology of the F1 tree suggested that *Citrus* and *Toona* formed a clade sister to the members of the Burseraceae (Fig. 2a), consistent with the species tree (Soltis et al., 2011); however, there was no support for this topological arrangement using F2 (Fig. 2b). Other clades in the "core TPSb" included clades with all paralogous copies of *Vitis*, *Solanum*, *Arabidopsis* and all the members of the Lamiaceae included in this study; this result was consistent with previous phylogenetic analyses of the TPS family (e.g., Dudareva et al., 2003; Chen et al., 2011). Relationships among these clades were poorly supported, but both gene trees were consistent with the species tree (Soltis et al., 2011). The only exception was the relationship of the TPSb copies in *Arabidopsis*, which formed a clade more closely related to taxa in the Gentianidae (*Solanum*, Lamiaceae) rather than to taxa within the Rosidae (e.g., *Quercus*, *Protium*) as would be expected in the species tree (Soltis et al., 2011). However, this topological arrangement was recovered with low support when data was analyzed as codon instead of nucleotides sequences (Fig. 3).

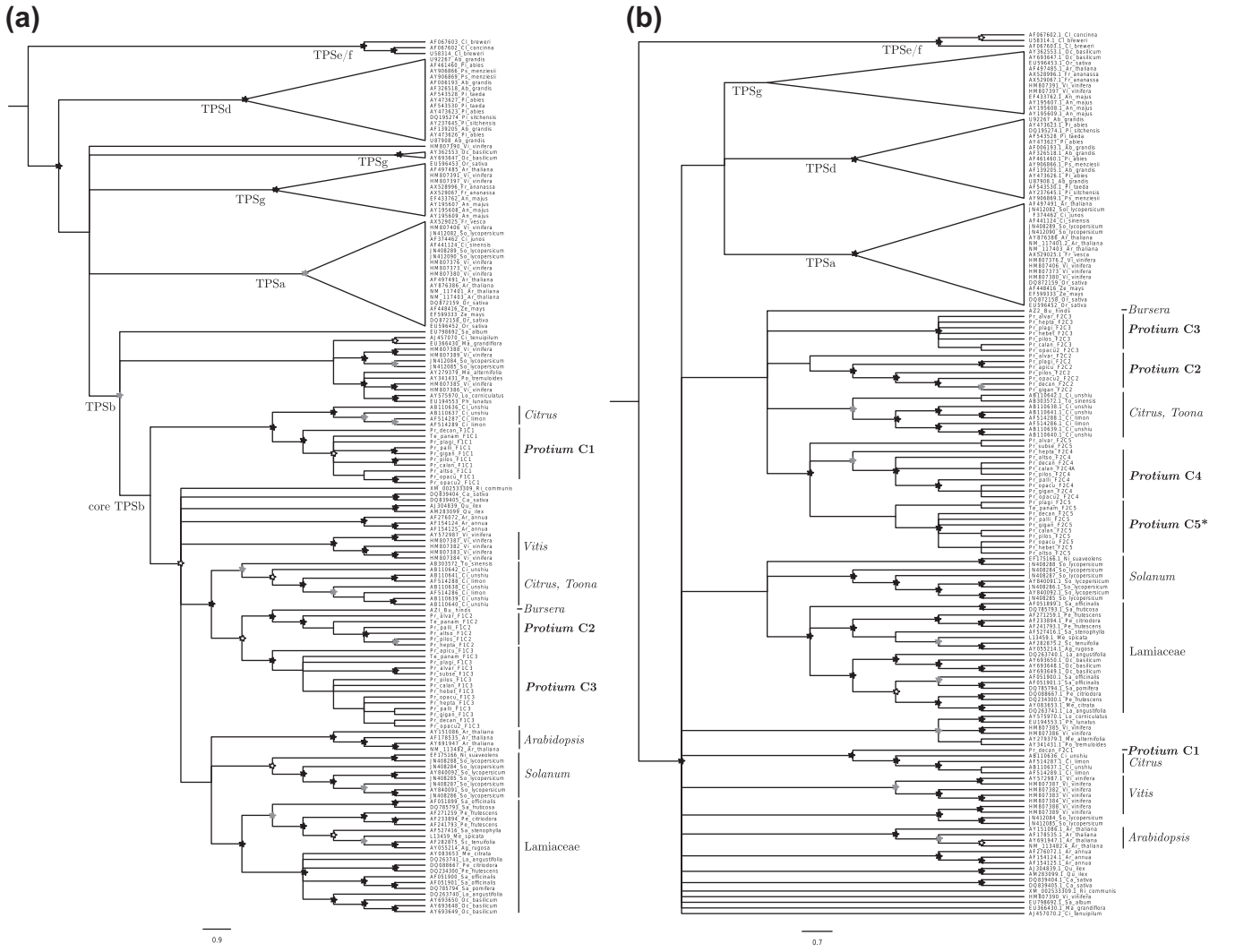

Fig. 3. Angiosperm-wide majority rule consensus trees from maximum likelihood-boostraped phylogenetic analyses of terpene synthase genes (TPS) using data as codons for (a) Fragment 1 and (b) Fragment 2. TPS subfamilies indicated under branches. Stars at each node represent bootstrap percentages (bp): 100% < bp ⩽ 90% (black); 90% < bp ⩽ 80% (gray); 80% < bp ⩽ 70% (white); bp < 70% (no symbol). Clades discussed in the text are indicated next to the tree with vertical lines. Gene Copy with signal of diversifying selection indicated with *.

3.3. Selection analyses

For sites models, maximum likelihood estimates of parameters and likelihood scores for all paralogous copies of F2 in *Protium* are presented in Table 3. Estimates of ω under models that allow for sites evolving under positive selection (M2a, M8) fit the data statistically better than the respective null models only for C5 (M2a vs. M1: LRT = 11.98, df = 2, P = 0.003; M8 vs. M7: LRT = 12.98, df = 2, P = 0.001). These findings suggested that about 2% of sites are evolving under diversifying selection, with ω around 7. Both models predicted with high posterior probability the same two sites evolving under positive selection at positions 137 and 157 (Fig. A1 of the Supplementary material). Substitutions in site 137 were due to changes in all three codon positions, and these changes alter the charge of the residue (Lys, Gly, or Ser). Substitutions in site 157 were due to changes at the second and third position of the codon resulting in residues with different polarities and physical properties (Thr, Arg, Ser, or Ile). M8 also fit the data statistically better than M7 for C2 (Table 3); however, no sites were predicted with high posterior probability to be evolving under positive selection.

A branch-sites model used to test for evidence of a burst of positive selection among sites along branches leading to all paralogous copies in *Protium* did not fit the data better than did the null model (LRT = 0.001, df = 1, P = 0.97) suggesting that any nonsynonymous changes along these branches may be the result of relaxed selection after duplication instead of diversifying selection.

4. Discussion

The production of monoterpenes is catalyzed by monoterpene synthases (TPSb), a diverse family of enzymes that catalyze the production of most monoterpene carbon skeletons (Davis and Croteau, 2000). In this study, we present the first phylogenetic study and molecular evolution analysis of TPSb in a diverse group of tropical trees, as a framework to investigate the patterns of sequence evolution that may be implicated in the evolution of chemical diversity in this genus, and shed light on the molecular mechanisms potentially driving such variation.

Our sequencing survey and phylogenetic analyses of TPSb reveal that *Protium* harbors at least three (Figs. 2a and 3a), and potentially five (Figs. 2b and 3b) copies of this gene subfamily. These analyses suggest that the duplication event giving rise to C1 occurred about 130–135 Myr (Smith et al., 2010), close to the common ancestor of the core Eudicots. Interestingly, this copy has apparently been maintained only in the genomes of *Protium* and *Citrus* – both members of the highly aromatic and chemically diverse clade Sapindales – and likely lost in all other angiosperms. However, the number of angiosperms for which the TPSb gene family has been well characterized is extremely low, thus this observation would require confirmation from studies in other

Table 3
Parameter estimates and likelihood scores under models of variable ω ratios among sites for each paralogous copy in F2 (TPSb genes) for *Protium*. The number after the model code, in parenthesis, is the number of free parameters. PSS is the number of positive selected sites inferred with the Bayes Empirical Bayes (BEB) approach.

Model	Copy	Parameter estimates	PSS	l
M1a: nearly neutral (2)	C2	p_0 = 0.7, (p_1 = 0.3) ω_0 = (0.12), (ω_1 = 1)		−1604.28
	C3	p_0 = 0.65, (p_1 = 0.35) ω_0 = (0.01), (ω_1 = 1)	NA	−1170.72
	C4	p_0 = 0.62, (p_1 = 0.38) ω_0 = (0), (ω_1 = 1)		−1106.85
	C5	p_0 = 0.62, (p_1 = 0.38) ω_0 = (0.10), (ω_1 = 1)		−1455.62
M2a: positive selection (4)	C2	p_0 = 0.80, p_1 = 0, p_2 = (0.2) (ω_0 = 0.18), (ω_1 = 1), ω_2 = 1.76		−1602.66
	C3	p_0 = 0.73, p_1 = 0.25, p_2 = (0.02) (ω_0 = 0.09), (ω_1 = 1), ω_2 = 68.9		−1168.51
	C4	p_0 = 0.7, p_1 = 0, p_2 = (0.3) (ω_0 = 0), (ω_1 = 1), ω_2 = 1.64		−1105.98
	C5	p_0 = 0.61, p_1 = 0.36, p_2 = (0.02) (ω_0 = 0.15), (ω_1 = 1), ω_2 = 6.96	2*	**−1449.63**
M7: β (2)	C2	p = 0.30, q = 0.46		−1606.23
	C3	p = 0.02, q = 0.04		−1170.76
	C4	p = 0.006, q = 0.009	NA	−1106.87
	C5	p = 0.10, q = 0.11		−1455.99
M8: β and ω (4)	C2	p_0 = 0.8, (p_1 = 0.2) p = 21.90, q = 99, ω = 1.77		**−1602.68**
	C3	p_0 = 0.98, (p_1 = 0.02) p = 0.21, q = 0.45, ω = 68.14		−1168.42
	C4	p_0 = 0.7, (p_1 = 0.3) p = 0.005, q = 2.65, ω = 1.64		−1105.98
	C5	p_0 = 0.97, (p_1 = 0.025) p = 0.35, q = 0.43, ω = 6.74	2**	**−1449.50**

Likelihood scores (l) in bold for statistically significant likelihood ratio test.
* if posterior probability > 0.95.
** if posterior probability > 0.99.

species. Nevertheless, earlier studies using sparser taxon sampling have noted that this copy in *Citrus* was only distantly related to other TPSb copies in this genus and to other angiosperms (Lücker et al., 2002; Shimada et al., 2004, 2005). Here, with an angiosperm-wide sampling including genomic-level sequences from *Arabidopsis*, *Solanum* and *Vitis*, we confirm that this copy is an ancient paralog, sister to the rest of the "core TPSb". Whether C1 is maintained in the genome of other taxa in the Sapindales is not known, but if so, this would imply selection for retention of functional copies over an extremely long period of time and could possibly be related to the high monoterpene diversity in this clade. It is unclear if copies C2, C3, C4, and C5 are restricted only to *Protium*, Burseraceae or the Sapindales. In either case, our results suggest that the duplication events giving rise to these copies occurred approximately during the last 50–70 Myr (Magallón and Castillo, 2009; De-Nova et al., 2012). Studies sampling TPSb genes from multiple species per family in the Sapindales – and in other angiosperms in general – are necessary to shed light into the extent of duplication events and the retention of different TPSb copies across genomes in the angiosperms.

That *Protium* harbors multiple copies of TPSb genes supports the hypothesis that the evolution of the monoterpene pathway, and thus the chemical profile diversity expressed in this genus, is positively related to gene copy number (Flagel and Wendel, 2009). This is consistent with previous analyses characterizing the TPS family in other angiosperms with high phenotypic diversity in terpenes (e.g., Martin et al., 2010). It is important to note, however, that we have not functionally characterized each copy in this study, therefore, a complete evaluation of this hypothesis will require experimental testing. It is likely that the three copies in F1 correspond to three copies in F2, and that we failed to recover the two extra copies in F1 that would correspond to the respective copies in F2. Conversely, given that the active site of the TPSb genes is located in the C-domain (i.e., F2) (Degenhardt et al., 2009), it is plausible that the N-domain (i.e., F1) is more conserved across closely related TPSb paralogs (under the assumption that each copy is functionally distinct), and thus different copies in F2 may share the same sequences for F1. A third possibility is that different paralogs may have alternative splicing variants (e.g., Keszei et al., 2010), but our approach precludes accurate assembling of all transcripts. In either case, future studies seeking to elucidate the evolutionary history of TPSb in *Protium* will benefit from further TPSb targeted sequencing, or from high throughput sequencing approaches to mine an mRNA library for this gene subfamily (e.g., Sun et al., 2010).

Although the overall topology of both phylogenetic trees is congruent with the angiosperm species tree (Soltis et al., 2011), the phylogenetic clustering of all paralogous copies per taxon is intriguing (Figs. 2 and 3). This pattern suggests either multiple duplication events with subsequent repeated losses throughout angiosperm evolution, or multiple more recent duplication events within species after species-level divergence. The first scenario requires postulating a very high number of duplications and losses, which is not parsimonious and rather improbable. By contrast, the second scenario is consistent with the tandem arrangement of several TPS gene copies on the chromosomes of different angiosperms (Chen et al., 2004; Martin et al., 2010), and thus provides a likely hypothesis for TPSb evolution; unfortunately, a chromosome map for *Protium* is not available. Alternatively, this phylogenetic pattern could also be explained by convergent sequence evolution on functional sites within species. This implies that functional sites maybe misleading for phylogenetic analyses and thus will tend to produce biased gene trees. To evaluate this hypothesis, we attempted to align nucleotide sites likely evolving neutrally in TPSb for *Protium* and *Arabidopsis*. It was not possible to generate a reliable alignment using either introns, third codon positions, or combining both classes of sites, suggesting that genes in this subfamily are evolving rapidly, and that our phylogenetic results using all codon positions may thus be representing an unbiased gene history.

Codon-based evolutionary analyses of the genes underlying the production of plant secondary compounds can provide valuable insights on the potential mechanisms underlying the diversification of plant defenses. For example, in *Arabidopsis*, Benderoth et al. (2006) showed that some methylthioalkylmalate (MAM) synthases, the central enzymes in the glucosinolate metabolism, are evolving under diversifying selection. Hydrolysis of glucosinolates generates biologically active compounds that play an important ecological role in plant defense against herbivorous insects. In contrast, similar codon-based analyses of the TPSb genes in *Protium* indicate that with the exception of C5 none of the other copies of TPSb are evolving under positive (diversifying) selection. This result is not consistent with the hypothesis that monoterpenes in *Protium* act as direct defenses against herbivores in a coevolutionary arms race-like scenario (Ehrlich and Raven, 1964). Rather, these results suggest that, on average, sites are evolving under a combination of stabilizing and relaxed selection, and thus it is likely that monoterpenes in this genus have been fine-tuned over millions of years for other functions such as the attraction of predators or parasitoids of herbivores (Mumm and Hilker, 2006; Schnee et al., 2006). In maize, two terpene synthases – TPS10, TPS23 – display a similar pattern of sequence evolution with stabilizing selection for the maintenance of function, indicating a prominent role of these enzymes in indirect defense against herbivores (Köllner et al., 2009). In rice, when different genetic lines are silenced for monoterpene volatile production, parasitoids are less attracted to mutant plants, parasitism on herbivore eggs is reduced by more than 50%, and population densities of predatory spiders are drastically reduced (Xiao et al., 2012). This suggests that the function of some monoterpenes may be considered an effective honest signal released by plants when attacked by herbivores, and thus selection should act to maintain this function. It is plausible that a similar mechanism operates to protect *Protium* trees regardless of the kind of herbivores that commonly attack them by attracting a diversity of parasitoids and predators. It is also plausible that monoterpenes in *Protium* may be part of a generic defense mechanism against enemies that evolved long ago in *Protium*'s common ancestor and it is shared by many extant species, thus monoterpene synthases today show evidence of stabilizing selection. Similarly, other functions of monoterpenes that would be beneficial to the plants and not be subjected to arms race type diversifying selection include communication cues between trees to alert the presence of enemies and initiate defense induction (Baldwin et al., 2006; Frost et al., 2007; Ton et al., 2006), or conferring protection to the photosynthetic machinery under thermal or oxidative conditions (Vickers et al., 2009). Furthermore, it seems likely that given the diversity of TPSb copies present in the genome of *Protium*, different genes may catalyze the synthesis of compounds with distinct functions that can help plants cope with the interaction of multiple biotic and abiotic factors (Holopainen and Gershenzon, 2010; Xiao et al., 2012). Alternatively, it is possible that some monoterpenes do function as direct defenses against herbivores, but even if herbivores are evolving effective counterdefenses in an arms race type way, selection on the plants is acting upstream or downstream in the biosynthesis of these compounds, either in other genes or at the level of gene regulation. Although little is known about this mechanism in the terpene biosynthetic pathway, changes in gene regulation have often been found in studies of biosynthetic pathways of other plant secondary compounds such as flavonoids (Koes et al., 2005) or aromatic amino acids (Tzin and Galili, 2010).

We did find evidence that sites 137 and 157 in C5 are evolving under positive selection (Table 3). Information from other plants

on the phenotypic effect of mutations in these specific positions is currently unavailable. However, position 157 occurs within a region that has been previously functionally characterized in a monoterpene synthase from *Salvia* (Kampranis et al., 2007). This region – referred to as region 2 in Kampranis et al. (2007) – comprises the sites 154–158 (Fig. A1 of the Supplementary material), which are part of the loop connecting helices α18 and α19 of the protein. Selective residue mutations in this region alter the 3D conformation and polarity of the catalytic site, which affect the rearrangements that the intermediate carbocations undergo during monoterpene synthesis (Davis and Croteau, 2000). Although the replacements occurring in site 157 in *Protium* are not shared with *Salvia*, it is possible that these mutations also have an effect on product specificity in C5 given the chemical properties (e.g., Ile) and sizes (e.g., Arg) of the alternative residues. Future site-directed mutagenesis studies can shed light on wild and mutant monoterpene profiles of these genes. Furthermore, functional and mutational analyses in other terpene synthases (e.g., diterpene synthases) demonstrate that changes in this region can be critical for catalysis (e.g., Keeling et al., 2008; Zerbe et al., 2012). This is consistent with the hypothesis that this region is likely an important element for functional diversification.

This study represents the first attempt to trace the molecular evolutionary history of the TPSb genes in *Protium*, a diverse clade of tropical trees. The primers we have developed proved successful at amplifying multiple paralogous copies in a diverse set of species. Therefore, it is now possible to use these primers and more sequences to assist the redesign of new specific primers for particular copies to study patterns of interspecific variation in deeper detail, or intraspecific variation across populations showing contrasting monoterpene profiles. We have shown that *Protium* retained at least three and maybe up to five copies of TPSb genes, and we suggest this may be associated with the monoterpene diversity expressed in this genus. Moreover, we have inferred that these genes appear to be evolving largely under relaxed or purifying selection, which suggests that, with the possible exception of C5, these genes may be involved in functions other than direct defenses against herbivores.

Acknowledgments

We especially thank Chelsea Specht for kindly providing access to her laboratory at UC Berkeley. We also thank Roxana Yockteng for help with RNA work and suggestions on data analysis. Jill Preston provided valuable insights into lab procedures, data analyses, and on earlier versions of this manuscript. Thanks to Santiago Ramírez for insightful discussions on plant–animal interactions and the molecular bases underlying such interactions, as well as for comments on the manuscript. Members of the Fine Lab provided valuable feedback during the development of this work. This work was supported by the National Science Foundation (DEB 0919567) and the UC Berkeley Hellman Fellows program.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2013.04.024.

References

Anisimova, M., Bielawski, J.P., Yang, Z.H., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18, 1585–1592.

Arimura, G.I., Huber, D.P.W., Bohlmann, J., 2004. Forest tent caterpillars (*Malacosoma disstria*) induce local and systematic emissions of terpenoid volatiles in hybrid poplar (*Populus trichocarpa x deltoides*): cDNA cloning, functional characterization, and patterns of gene expression of (−)-germacrene D synthase, PtdTPS1. Plant J. 37, 603–616.

Aubourg, S., Lecharny, A., Bohlmann, J., 2002. Genomic analysis of the terpenoid synthases (AtTPS) gene family of *Arabidopsis thaliana*. Mol. Genet. Genomics 267, 730–745.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–141.

Bohlmann, J., Meyer-Gauen, G., Croteau, R., 1998. Plant terpenoid synthases: molecular biology and phylogenetic analysis. Proc. Natl. Acad. Sci. USA 95, 4126–4133.

Baldwin, I.T., Halitschke, R., Paschold, A., von Dahl, C.C., Preston, C.A., 2006. Volatile signaling in plant–pant interactions: "Talking trees" in the genomic era. Science 311, 812–815.

Benderoth, M., Textor, S., Windsor, A.J., Michell-Olds, T., Gershenzon, J., Kryomann, J., 2006. Positive selection driving diversification in plant secondary metabolism. Proc. Natl. Acad. Sci. USA 103, 9118–9123.

Bernhardsson, C., Ingvarsson, P.K., 2012. Geographical structure and adaptive populations differentiation in herbivore defence genes in European aspen (*Populus tremula* L., Salicaceae). Mol. Ecol. 21, 2197–2207.

Chen, F., Ro, D.K., Petri, J., Gershenzon, J., Bohlmann, J., Pichersky, E., Tholl, D., 2004. Characterization of a root-specific *Arabidopsis* terpene synthase responsible for the formation of the volatile monoterpene 1,8-cineole. Plant Physiol. 135, 1956–1966.

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.

Daly, D.C., 2007. A new section of *Protium* Burm. f. from the neotropics. Stud. Neotrop. Burseraceae XIV. Brittonia 59, 1–24.

Daly, D.C., Fine, P.V.A., 2011. A new Amazonian section of *Protium* (Burseraceae) including both edaphic specialist and generalist taxa. Studies in neotropical Burseraceae XVI. Syst. Bot. 36, 939–949.

Daly, D.C., Fine, P.V.A., Martinez-Habibe, C.M., 2012. Burseraceae: a model for studying the Amazonian flora. Rodriguesia 63, 21–30.

Davis, E., Croteau, R., 2000. Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. Top. Curr. Chem. 209, 53–95.

Degenhardt, J., Köllner, T.G., Gershenzon, J., 2009. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. Phytochemistry 70, 1621–1637.

de Freitas, J.G.R., da Camara, C.A.G., de Moraes, M.M., da Silva, H.C.H., 2011. Volatile constituents of two species of *Protium* from the Atlantic rainforest in the state of Pernambuco, Brazil. Nat. Prod. Commun. 6, 1727–1730.

De-Nova, J.A., Medina, R., Montero, J.C., Weeks, A., Rosell, J.A., Olson, M.E., Eguiarte, L.E., Magallón, S., 2012. Insights into the historical construction of species-rich Mesoamerican seasonally dry tropical forests: the diversification of *Bursera* (Burseraceae, Sapindales). New Phytol. 193, 276–287.

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wislon, A., 2011. Geneious v5.5.6. <http://www.geneious.com>.

Dudareva, N., Martin, D., Kish, C.M., Kolosova, N., Gorenstein, N., Faldt, J., Miller, B., Bohlmann, J., 2003. (E)-β-ocimene and myrcene synthases genes of floral scent biosynthesis in snapdragon: function and expression of three terpene synthase genes of a new terpene synthase subfamily. Plant Cell 15, 1227–1241.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Ehrlich, P., Raven, P., 1964. Butterflies and plants: a study in coevolution. Evolution 18, 586–608.

El Tamer, M.K., Lücker, J., Bosch, D., Verhoeven, H.A., Verstappen, F.W.A., Schwab, W., van Tunen, A.J., Voragen, A.G.J., de Maagd, R.A., Bouwmeester, H.J., 2003. Domain swapping of *Citrus limon* monoterpene synthases: impact on enzymatic activity and product specificity. Arch. Biochem. Biophys. 411, 196–203.

Evans, P.H., Becerra, J.X., Venable, D.L., Bowers, W.S., 2000. Chemical analysis of squirt-gun defense in *Bursera* and counterdefense by chrysomelid beetles. J. Chem. Ecol. 26, 645–754.

Falara, V., Akhtar, T.A., Nguyen, T.T.H., Spyropoulou, E.A., Bleeker, P.M., Schauvinhold, I., Matsuba, Y., Bonini, M.E., Schilmiller, A.L., Last, R.L., Schuurink, R.C., Pichersky, E., 2011. The tomato terpene synthase gene family. Plant Physiol. 157, 770–789.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Fine, P.V.A., Daly, D.C., Villa Muñoz, G., Mesones, I., Cameron, K.M., 2005. The contribution of edaphic heterogeneity to the evolution and diversity of Burseraceae trees in the western Amazon. Evolution 59, 1464–1478.

Fine, P.V.A., Miller, Z.J., Mesones, I., Irazuzta, S., Appel, H.M., Stevens, M.H.H., Slååksjårvi, I., Schultz, J.C., Coley, P.D., 2006. The growth-defense trade-off and habitat specialization by plant in Amazonian forests. Ecology 87, S150–S162.

Flagel, L.E., Wendel, J.F., 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183, 557–564.

Frost, C.J., Appel, H.M., Carlson, J.E., De Moraes, C.M., Mescher, M.C., Schultz, J.C., 2007. Within-plant signalling via volatiles overcomes vascular constraints on systemic signalling and primes responses against herbivores. Ecol. Lett. 10, 490–498.

Gershenzon, J., Dudareva, N., 2007. The function of terpene natural products in the natural world. Nat. Chem. Biol. 3, 408–414.

Goldman, N., Yang, Z.H., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Guoy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequencing alignments and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.
Hampel, D., Mosandl, A., Wüst, M., 2006. Biosynthesis of mono- and sesquiterpenes in strawberry fruits and foliage: 2H labeling studies. J. Agric. Food Chem. 54, 1473–1478.
Holopainen, J.K., Gershenzon, J., 2010. Multiple stress factors and the emission of plants VOCs. Trends Plant Sci. 15, 176–184.
Kampranis, S.C., Ioannidis, D., Purvis, A., Mahrez, A., Mahrez, W., Ninga, E., Katerelos, N.A., Anssour, S., Dunwell, J.M., Degenhardt, J., Makris, A.M., Goodenough, P.W., Johnson, C.B., 2007. Rational conversion of substrate and product specificity in a *Salvia* monoterpene synthase: structural insights into the evolution of terpene synthase function. Plant Cell 19, 1994–2005.
Katoh, K., Asimenos, G., Toh, H., 2009. Multiple alignment of DNA sequences with MAFFT. In: Posada, D. (Ed.), Bioinformatics for DNA Sequence Analysis, Methods in Molecular Biology, vol. 537. Springer Science: Human Press, pp. 39–64.
Keeling, C.I., Bohlmann, J., 2006. Genes, enzymes and chemical of terpenoid diversity in the constitutive and induced defence of conifers against insect and pathogens. New Phytol. 170, 657–675.
Keeling, C.I., Weisshaar, S., Lin, R.P.C., Bohlmann, J., 2008. Functional plasticity of paralogous diterpene synthases involved in conifer defense. Proc. Natl. Acad. Sci. USA 105, 1085–1090.
Keeling, C.I., Weisshaar, S., Ralph, S.G., Jancsik, S., Hamberger, B., Dulalt, H.K., Bohlmann, J., 2011. Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*Picea* spp.). BMC Plant Biol. 11, 43.
Keszei, A., Brubaker, C.L., Foley, W.J., 2008. A molecular perspective on terpene variation in Australian Myrtaceae. Aust. J. Bot. 56, 197–213.
Keszei, A., Brubaker, C.L., Carter, R., Köllner, T., Degenhardt, J., Foley, W.J., 2010. Functional and evolutionary relationships between terpene synthases from Australian Myrtaceae. Phytochemistry 71, 844–852.
Koes, R., Vermeij, W., Quattrocchio, F., 2005. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci. 10, 236–242.
Köllner, T.G., Gershenzon, J., Degenhardt, J., 2009. Molecular and biochemical evolution of maize terpene synthase 10, an enzyme of indirect defense. Phytochemistry 70, 139–1145.
Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Langenheim, J.H., 2003. Pant resins: chemistry, Evolution, Ecology, Ethnobotany. Timber Press, Portland OR, USA.
Lima Filho, D.A., Matos, F.D.A., Amaral, I.L., Revilla, J., Coêlho, L.S., Ramos, J.F., Santos, J.L., 2001. Inventário florístico de floresta ombrófila densa de terra firme na região do Rio Urucu-Amazonas, Brasil. Acta Amazônica 31, 565–579.
Lücker, J., El Tamer, M., Schwab, W., Verstappen, F.W.A., van der Plas, L.H.W., Bouwmeester, H.J., Verhoeven, H.A., 2002. Monoterpene biosynthesis in lemon (*Citrus limon*). Eur. J. Biochem. 269, 3160–3171.
Magallón, S., Castillo, A., 2009. Angiosperm diversification through time. Am. J. Bot. 96, 349–365.
Marques, D.D., Sartori, R.A., Gomes Lemos, T.L., Machado, L.L., Nery de Souza, J.S., Queiroz Monte, F.J., 2010. Chemical composition of the essential oils from two subspecies of *Protium heptaphyllum*. Acta Amazonica 40, 227–230.
Martin, D.M., Aubourg, S., Schouweny, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T., Bohlmann, J., 2010. Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on a genome assembly, FLcDNA cloning, and enzyme assays. BMC Plant Biol. 10, 226.
Matos, F.D.A., Amaral, I.L., 1999. Análise ecologica de um hectare em floresta ombrófila densa de terra-firme, Estrada da Várzea, Amazônas, Brasil. Acta Amazônca 29, 365–379.
Mumm, R., Hilker, M., 2006. Direct and indirect chemical defence of pine against folivorous insects. Trends Plant Sci. 11, 351–358.
Pearson, W.R., Lipman, D.J., 1998. Improved tool for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.
Pearson, W.R., Wood, T.C., 2007. Statistical significance in biological sequence comparison. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics, third ed. John Wiley and Sons, England, pp. 40–66.
Peters, R.J., Croteau, R.B., 2003. Alternative termination chemistries utilized by monoterpene cyclases: chimeric analysis of bornyl diphophate, 1,8-cineole, and sabine synthases. Arch. Biochem. Biophys. 417, 203–211.
Pichersky, E., Gang, D.R., 2000. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci. 5, 439–445.
Pichersky, E., Gershenzon, J., 2002. The formation and function of plant volatiles: perfumes for pollinator attraction and defense. Curr. Opin. Plant Biol. 5, 237–243.
Rankin-de-Morona, J.M., Prance, G.T., Hutching, R.W., Silva, M.F., Rodrigues, W.A., Uehling, M.E., 1992. Preliminary results of a large scale tree inventory of upland rain forest in the central Amazon. Acta Amazônica 22, 493–534.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Schnee, C., Köllner, T.G., Held, M., Turlings, T.C.J., Gershenzon, J., Degenhardt, J., 2006. The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. Proc. Natl. Acad. Sci. USA 103, 1129–1134.
Sharkey, T.D., Yeh, S., Wiberley, A.E., Falbel, T.G., Gong, D., Fernandez, D.E., 2005. Evolution of the isoprene biosynthetic pathway in kudzu. New Phytol. 137, 700–712.
Shimada, T., Endo, T., Fujii, H., Hara, M., Ueda, T., Kita, M., Omura, M., 2004. Molecular cloning and functional characterization of four monoterpene synthase genes from *Citrus unshiu* Marc. Plant Sci. 166, 49–58.
Shimada, T., Endo, T., Fujii, H., Hara, M., Omura, M., 2005. Isolation and characterization of (E)-beta-ocimene and 1,8 cineole synthases in Citrus unshiu Marc. Plant Sci. 168, 987–995.
Shishova, E.Y., Di Costanzo, L., Cane, D.E., Christianson, D.W., 2007. X-ray crystal structure of aristolochene synthase from *Aspergiullus terreus* and the evolution of templates for the cyclization of farnesyl diphosphate. Biochemistry 46, 1941–1951.
Siani, A.C., Garrido, I.S., Monteiro, S.S., Carvalho, E.S., Ramos, M.F.S., 2004. *Protium icicariba* as a resource of volatile essences. Biochem. Syst. Ecol. 32, 477–489.
Silva, J.R.deA., Zoghbi, M.dasG.B., Pinto, A.daC., Godoy, R.I.O., Amaral, A.C., 2009. Analysis of the hexane extracts from seven oleoresins of Protium species. J. Essent. Oil Res. 21, 305–308.
Smith, S.A., Beaulieu, J.M., Donoghue, M.J., 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl. Acad. Sci. USA 107, 5897–5902.
Snoeren, T.A.L., Kappers, I.F., Broekgaarde, C., Mumm, R., Dicke, M., Bouwmeester, H.J., 2010. Natural variation in herbivore-induced volatiles in *Arabidopsis thaliana*. J. Exp. Bot. 61, 3041–3056.
Soltis, D.E., Smith, S.A., Cellinese, N., Wurdack, K.J., Tank, D.C., Brockington, S.F., Refulio-Rodriguez, N.F., Walker, J.B., Moore, M.J., Carlsward, B.S., Bell, C.D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C.A., Gitzendanner, M.A., Systma, K.J., Qiu, Y.L., Hilu, K.W., David, C.D., Sanderson, M.J., Beaman, R.S., Olmstead, R.G., Judd, W.S., Donoghue, M.J., Soltis, P.S., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98, 704–730.
Sullivan, J., Joyce, P., 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36, 445–466.
Sun, C., Li, Y., Wu, Q., Luo, H., Sun, Y., Song, J., Lui, E.M.K., Chen, S., 2010. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11, 262.
Ton, J., D'Alessandro, M., Jourdine, V., Jakab, G., Karlen, D., Held, M., Mauch-Mani, B., Turlings, T.C.J., 2006. Priming by airborne signals boosts direct and indirect resistance in maize. Plant J. 49, 16–26.
Trapp, S.C., Croteau, R.B., 2001a. Defensive resin biosynthesis in conifers. Annu. Rev. Plant Phys. 52, 689–724.
Trapp, S.C., Croteau, R.B., 2001b. Genomic organization of plant terpene synthases and molecular evolutionary implications. Genetics 158, 811–832.
Tzin, V., Galili, G., 2010. New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. Mol. Plant 3, 956–972.
Vickers, C.E., Gershenzon, J., Lerdau, M.T., Loreto, F., 2009. A unified mechanism of action for volatile isoprenoid in plant abiotic stress. Nat. Chem. Biol. 5, 283–291.
Weeks, A., Daly, D.C., Simpson, B.B., 2005. The phylogenetic history and biogeography of the frankincense and myrrh family (Burseraceae) based on nuclear and chloroplast sequence data. Mol. Phylogenet. Evol. 35, 85–101.
Whittington, D.A., Wise, M.L., Urbansky, M., Coates, R.M., Croteau, R.B., Christianson, D.W., 2002. Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. Proc. Natl. Acad. Sci. USA 99, 15375–15380.
Wong, W., Yang, Z., Goldman, N., Nielsen, R., 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168, 1041–1051.
Xiao, Y., Wang, Q., Erb, M., Turlings, T.C.J., Ge, L., Hu, L., Li, J., Han, X., Zhang, T., Lu, J., Zhang, G., Lou, Y., 2012. Specific herbivore-induced volatiles defend plants and determine insect community composition in the field. Ecol. Lett.. http://dx.doi.org/10.1111/j.1461-0248.2012.01835.x.
Yang, Z., 2007. PAML4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15, 496–503.
Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.
Yang, Z., Wong, W., Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.
Zhang, J., Nielsen, R., Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479.
Zerbe, P., Chiang, A., Bohlmann, J., 2012. 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. Phytochemsitry 74, 30–39.
Zwickl, D.J., 2006. Genetic Algorithm Approaches for Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. PhD Thesis, The University of Texas at Austin, Austin, TX, USA.