genes

MDPI

Article

Leaf Transcriptome Assembly of *Protium copal* (Burseraceae) and Annotation of Terpene Biosynthetic Genes

Gabriel Damasco [1,*], Vikram S. Shivakumar [1], Tracy M. Misciewicz [2], Douglas C. Daly [3] and Paul V. A. Fine [1]

[1] Department of Integrative Biology and University and Jepson Herbaria, University of California, Berkeley, CA 94720, USA; vshiv@berkeley.edu (V.S.S.); paulfine@berkeley.edu (P.V.A.F.)
[2] Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019, USA; tmisiewicz@ou.edu
[3] Institute of Systematic Botany, The New York Botanical Garden, Bronx, NY 10458, USA; ddaly@nybg.org
* Correspondence: gdamasco@berkeley.edu

Received: 7 March 2019; Accepted: 20 May 2019; Published: 22 May 2019

Abstract: Plants in the Burseraceae are globally recognized for producing resins and essential oils with medicinal properties and have economic value. In addition, most of the aromatic and non-aromatic components of Burseraceae resins are derived from a variety of terpene and terpenoid chemicals. Although terpene genes have been identified in model plant crops (e.g., *Citrus*, *Arabidopsis*), very few genomic resources are available for non-model groups, including the highly diverse Burseraceae family. Here we report the assembly of a leaf transcriptome of *Protium copal*, an aromatic tree that has a large distribution in Central America, describe the functional annotation of putative terpene biosynthetic genes and compare terpene biosynthetic genes found in *P. copal* with those identified in other Burseraceae taxa. The genomic resources of *Protium copal* can be used to generate novel sequencing markers for population genetics and comparative phylogenetic studies, and to investigate the diversity and evolution of terpene genes in the Burseraceae.

Keywords: RNA-sequencing; de novo assembly; *Protium*; terpenoid; copal resin

1. Introduction

The Burseraceae family harbors over 600 species of plants, many of which produce resins and essential oils that have been economically and medicinally important for millennia in the Neotropics as well as parts of Asia and Africa. Uses of myrrh (*Commiphora* spp.) and frankincense (*Boswellia* spp.) are mentioned in biblical texts, and they are still used today in many religious rituals [1]. Mayan records from South and Central America dating back to 600 BC describe the use of copal (*Protium* sp. and *Bursera* sp.) resins as incense and medicines [2]. Today, Burseraceae plant species continue to be recognized globally for their medical, aromatic, flammable, and adhesive properties.

The majority of resins and oils produced by Burseraceae plants are terpenes. Most of these are produced internally by plant secretory cells and are defined as aromatic soluble mixtures of volatiles and non-volatiles rich in terpenoids. While the production of hundreds of different terpene chemicals occurs through ancient biosynthetic pathways common among all plants, different lineages are also known to produce their own specialized terpenes. Currently, the number of described terpenes already exceeds 60,000 [3], and as new specialized terpenes are discovered and described, that number will continue to increase [4]. In highly diverse plant families such as Burseraceae, terpenes include monoterpenes, sequiterpenes, diterpenes, and triterpenes found in the shoots, leaves, and flowers [1].

Terpene chemicals play major biological roles in a plant's metabolism and life cycle. For instance, primary terpenoids are major constituents of plant membranes and are important for maintaining the fluidity of these cells [5]. Defense against natural enemies (i.e., herbivores, pathogens, viruses) is another key function of terpenoids, whether directly targeting enemies as toxins and repellents, or indirectly through the attraction of predators or parasitoids of such enemies [6]. Although the functional aspects of terpene chemicals as an important defense strategy in plants is well studied, understanding the evolutionary origin of many different kinds of specialized terpenes remains a key goal.

Protium, a highly diverse genus within Burseraceae, has been an important non-model system to study the evolution of terpene metabolites. Some of the few available studies of terpene diversification have focused on *Protium* [7] and have found that coevolutionary relationships with natural enemies have played an important role in shaping the chemical diversity of *Protium* species [8]. It has also been hypothesized that the intrinsic genetic variability that encodes the production of terpene metabolites could also facilitate the diversity of terpenoid structures in plant lineages [5]. Despite the extensive knowledge of *Protium* chemistry [9], the diversity of genes responsible for synthesizing the array of terpenoids found in the genus is unknown and few genomic resources are available [7].

Here, we report the transcriptome assembly and annotation of terpene biosynthetic genes in *Protium copal* (Schltdl. & Cham.) Engl., a widespread subcanopy tree spanning environments from moist evergreen forests to seasonally deciduous forest in Central America. This novel genomic resource is now available, and several genetic markers that we expect to be useful for probe design are provided. In addition, we compare the diversity of terpene genes identified in *P. copal* to genes found in two closely related Burseraceae species (*Boswellia sacra* Flueck. and *Bursera simaruba* (L.) Sarg.). It has been well established that plants in the tropics experience much higher levels of herbivory and pathogen attack compared to other biomes [10]. Thus, we predict that: *P. copal* and *B. simaruba,* both found across Central America will have a higher diversity of terpenoid genes as compared to *B. sacra* which is endemic to the Arabian Peninsula and Northern Africa.

2. Materials and Methods

2.1. Study System

Protium, a genus of understory, midcanopy and canopy trees up to 35 m (115 ft.) tall, is one of the most speciose genera in the Burseraceae family. [11]. Protium is almost exclusively dioecious and fruits are dispersed by birds and arboreal mammals [11]. Protium copal, commonly known as the copal tree, is endemic to Mexico and Central America and it can grow up to 30 m (98 ft.) in height. While no estimates of genome size exist for *P. copal*, DNA C-values have been calculated for another Protium species, Protium serratum (Wall. ex Colebr.) Engl., which is reported to have a genome size of about 900 Mbp (http://data.kew.org/cvalues).

2.2. Leaf Material and RNA Isolation

Mature leaves were harvested from one cultivated specimen of *Protium copal* (see voucher details in GenBank: KJ503745 accession) and stored in liquid nitrogen prior to processing. The specimen accession is housed at the New York Botanical Garden. RNA isolation was performed as follows: The leaf sample was mixed with liquid nitrogen and ground into a fine powder using mortar and pestle. Total RNAs were extracted from approximately 100 mg of powdered tissue using a Qiagen RNeasy plant mini kit. cDNA libraries were constructed for paired-end 2x100 bp HiSeq 3000 platform (Illumina, San Diego, CA, USA). Total RNA concentration was assessed with a Qubit® 2.0 Fluorometer and purity was assessed using NanoDrop-ND 2000C spectrophotometer and bioanalyzer. The sample selected for sequencing had RNA integrity number (RINe) greater than 8.0, nanodrop ratios of 1.9–2.1 (260/230) and ratios of 2.0–2.5 (260/280). Library preparation and sequencing were performed at the QB3 Vincent J. Coates Genomics Sequencing Laboratory, University of California, Berkeley.

2.3. Transcriptome De Novo Assembly

Raw reads were filtered using quality value (Q) ≥ 30 and demultiplexed using an option of one mismatch in the index. Sequence quality was observed using FastQC 0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired-end reads were trimmed to remove Illumina sequencing adaptors, and poor-quality portions of reads using the default settings in Trimmomatic [12]. We used Trinity [13] version 2.5.1 with default parameters and a minimum contig length of 200 bp for assembly generation. Approximately 182 million paired-end reads were used to generate the de novo assembly. Transcript abundance was analyzed using Kallisto version 0.43.1 [14], an alignment-free abundance estimation tool that uses pseudo-alignment to calculate a transcript per million (TPM) value for each transcript. Transcript abundance was used to determine the abundance of terpenoid pathway transcripts. Transcriptome completeness was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) [15], which searched our transcriptome for 2,121 single-copy orthologs conserved among eudicots. We compared the exN50 and BUSCO results for *P. copal* with those of *B. sacra* and *B. simaruba* to understand how these transcriptomes differ in assembly quality.

2.4. Functional Annotation

Functional annotation of the *Protium copal* transcriptome was carried out using the Trinotate pipeline (https://github.com/Trinotate/Trinotate) as follows. TransDecoder v2.0.1 [16] was used to identify open reading frames (ORFs) and predict potential coding transcripts. The retrieved nucleotide sequences and putative protein sequences were then functionally annotated by using BLASTx and BLASTp to search for homologies (e-value $< 1 \times 10^{-10}$) against known protein and nucleotide sequences in the the UniProtKB/Swiss-Prot databases. Protein domains were identified using in the Pfam domain database (https://pfam.xfam.org/) using HMMER v3.1b2 [17]. The maximum e-value for reporting the best hit and associated annotation was 1×10^{-5}. EggNog and Kegg databases, which organize proteins based on orthology, were used to predict the function of our putative genes [18]. Blast2GO v3.1 [19] was also used to detect GO terms associated with biological processes (BPs), molecular functions (MFs), and cellular components (CCs) [19,20]. Transcripts were searched against the non-redundant (nr) database using BLASTx, which uses the translated protein sequence of each transcript as the search query [21]. The sequence description from the top homologous BLAST hit (e-value $< 10^{-5}$) was transferred to each transcript.

2.5. Phylogenetic Validation of Terpene Synthase Gene Annotations

P. copal transcripts identified as putative terpene synthase (TPS) genes were further validated using phylogenetic analysis. Protein sequences from known TPS genes from *Arabidopsis thaliana*, *Vitis vinifera*, *Rincus communis*, *Solanum lycopersicum*, *Medicago truncatula*, *Citrus junos*, *Citrus cinensis*, *Citrus unshiu*, *Citrus limon* were downloaded from GenBank and UniProt. Gymnosperm TPSd sequences were also downloaded from UniProt for use as outgroups. Nucleotide sequences from the *Protium copal* transcriptome were translated into amino acids for all frames using ORFinder (https://www.ncbi.nlm.nih.gov/orffinder/). The longest ORF that also blasted to known TPS genes using SmartBLAST were selected for inclusion in the final alignment. Nucleotide sequences were aligned using MAFFT version 7 with the E-INS-i algorithm using the MAFFT online service [22,23]. Final alignments are provided in the Supplementary Material. Maximum likelihood analyses were performed using the RAxML-HPC2 Workflow on XSEDE version 8.2.9 tool on Cipres (CyberInfrastructure for Phylogenetic RESearch) [24]. Bootstrap analysis was performed with 500 replicates using the Protein CAT model. Default parameters were employed for all other settings. The resulting phylogeny was visualized using Interactive Tree of Life (iTOL) [25].

2.6. *Comparison of Orthologous Clusters among* P. copal *and Closely Related Species*

Trinotate annotations were used to compare the diversity of orthologous gene clusters involved in terpene biosynthesis for *P. copal*, *B. sacra* and *B. simaruba*. Transcripts with sequences related to terpene synthase activity and terpenoid biosynthetic process were selected using the seqtk toolkit (https://github.com/lh3/seqtk). These terpene related transcripts were then translated to their corresponding protein sequence using the transeq function in EMBOSS version 6.6.0 (http://emboss.open-bio.org/rel/dev/apps/transeq.html). Orthologous gene prediction was performed using OrthoVenn (http://www.bioinfogenome.net/OrthoVenn/index.php). OrthoVenn performs an all-against-all BLASTP alignment, identifies putative orthology and generates disjoint clusters of closely related proteins using a Markov Clustering Algorithm (MCL) [26].

2.7. *Identification of Single Sequence Repeat (SSR) Markers and Transcripts with Single Nucleotide Polymorphisms (SNPs)*

Protium copal leaf transcripts were scanned for simple sequence repeat (SSR) markers using MISA version 1.0 [27]. The minimum number of repeat units was defined as 10 units for mono-nucleotide repeats, 6 units for di-nucleotide, and 5 units for tri-, tetra-, penta-, and hexa-nucleotide repeats. The maximum distance between two separate repeat regions was set at 100 bp. In addition, SNPs were discovered in the assembled transcriptome using Kissplice version 2.4.0-p1 [28], which analyzes RNA-seq reads to identify single nucleotide polymorphism in the sequencing library. Kissplice identifies variants within the transcriptome by aligning the sequence reads back to the assembled transcriptome in order to find variation within each individual transcript. This pipeline identifies variants that may be phylogenetically informative for designing putative orthologous markers for next-generation sequencing. The specific identified variants are also those for which this individual is heterozygous. To get positional data for the discovered SNPs, Transdecoder version 5.0.2 [16] was used to identify the open ORFs of each transcript, and BLAT version 36x2 [29] was used to align the identified SNPs to the transcriptome assembly, all using the default parameters.

3. Results

3.1. *Leaf Transcriptome Sequencing and De Novo Assembly*

Illumina sequencing generated 140 GB of data containing ca. 182 million paired-end reads. Trinity assembly resulted in 78,807 transcripts with fragment sizes ranging from 224 to 12,395 bp (Figure 1a) and a median contig length of 595 bp. Table 1 presents assembly statistics of the transcriptome, including transcripts that were retained with an exN50 length of 1145 bp, which represents the shortest transcript length at which half the assembled base pairs can be found. The exN50 statistic for the *P. copal* suggests the Trinity assembly is of high quality.

Table 1. Trinity transcriptome assembly statistics for *Protium copal* (Burseraceae).

Total trinity genes	**44,754**
Total trinity transcripts	63,288
Percent GC	41.56
Contig exN10	2,897
Contig exN20	2,102
Contig exN30	1,696
Contig exN40	1,396
Contig exN50	1,145
Median contig length:	595
Average contig	832.27
Total assembled bases	52,672,403

Furthermore, comparison of BUSCO analyses for all three transcriptomes (*P. copal*, *B. sacra* and *B. simaruba*) suggest that the *P. copal* assembly is the highest quality of the three with slightly better scores than *B. sacra* and significantly better scores than *B. simaruba* (Figure 1b). Of the 2121 single-copy orthologs included in the BUSCO analysis, 1292 (60.9%) were found to be complete, of which 563 (26.5%) were duplicated. Five hundred and fifteen orthologs (24.3%) were fragmented, and 314 (14.8%) were missing. Of the three Burseraceae transcriptomes, *Protium copal* was the most complete, with the fewest missing and fragmented orthologs (Figure 1c).

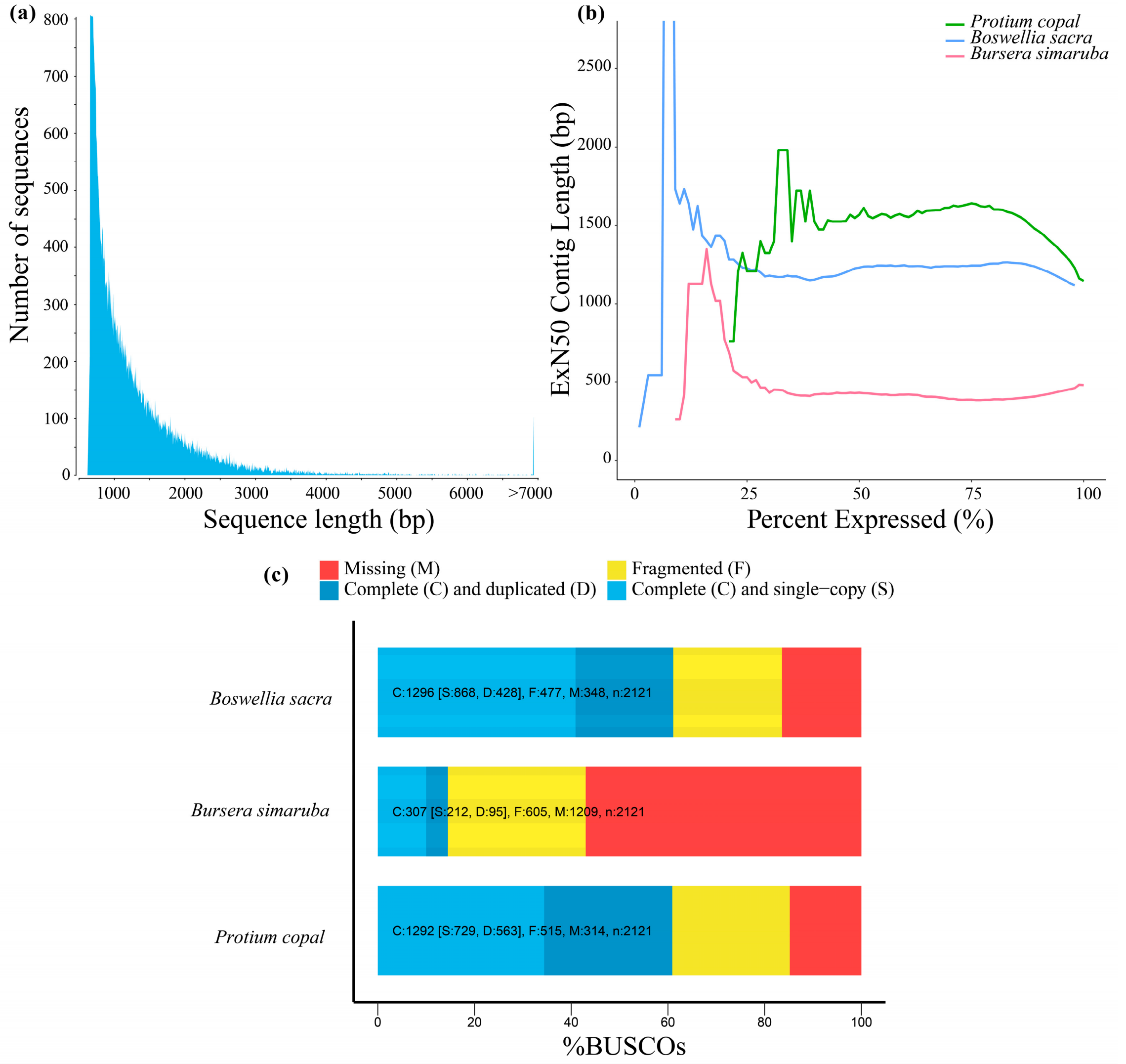

Figure 1. (**a**) Diagram of sequence length for the transcriptome assembly of *Protium copal* (Burseraceae). (**b**) The N50 value, the shortest transcript length at which 50% of assembled bases can be found, calculated only for the top percentile of expressed transcripts. The maximum exN50 value is at 75% expression, with an N50 of 1639. (**c**) Benchmarking Universal Single-Copy Orthologs (BUSCO) results of the *Protium copal* transcriptome assembly in comparison to two other transcriptomes in Burseraceae (*Bursera simaruba* and *Boswellia sacra*).

3.2. Functional Annotation and Similarity with Other Plant Genomes

The Trinotate annotation identified 38,042 nucleotide sequences (48.27%) and 24,250 protein sequences (30.77%) that showed significant homology with Viridiplantae when compared to the UniProtKB/Swiss-Prot database using BLASTx and BLASTp searches, respectively. Furthermore, 20,271 (25.72%) unique Pfam protein motifs were assigned (Table S1). These protein domains can be involved in various biological processes and molecular function such as protein–protein interactions, transcription regulation, and organic compound biosynthetic processes. In contrast, Blast2GO annotation retrieved 48,951 transcripts (69.91%) were successfully matched to homologous sequences for Viridiplantae in the NR database (e-value $< 1 \times 10^{-5}$). Out of the blasted transcripts, 18,822 transcripts (26.87%) were mapped to at least one gene ontology (GO) term using Blast2GO. The vast majority of BLAST hits for the transcripts came from the genus *Citrus*, also a member of the order Sapindales (Figure 2a). Our transcript dataset displayed 26,223 (37.44%) hits with the NR database and to *Citrus sinensis*, followed by *Citrus clementina* (10,185 top-hits, 14.54%), *Theobroma cacao* (2010 top-hits), and *Hevea brasiliense* (1623 top-hits) and *Vitis vinifera* (1484 top-hits) (Figure 2A).

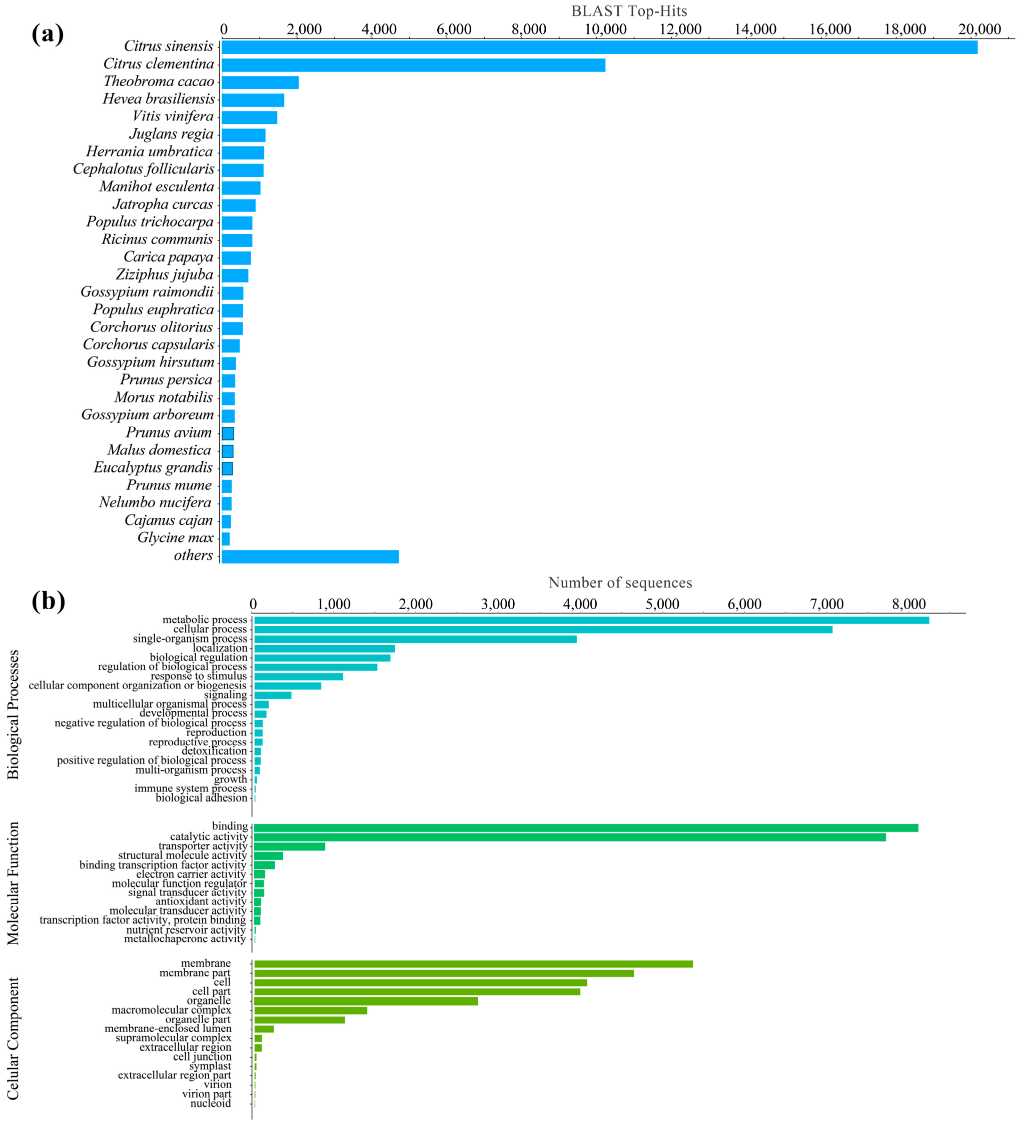

Figure 2. (**a**) The species distribution of BLAST hits used to assign sequence descriptions and Gene Ontology terms to transcripts for annotation. The majority of transcripts had BLAST hits from the genus Citrus, a member of the same order Sapindales. (**b**) Top 20 Gene Ontology terms (Level 2) for each of the three main sub-categories, Molecular Function (MF), Biological Process (BP), and Cellular Component (CC), based on the number of transcripts assigned that GO term.

3.3. Gene Ontology (GO) and Cluster of Orthologous Groups (COG)

GO analysis revealed 4532 unique GO terms related to plant gene ontology. Among the three main categories, Biological processes (BP) category was the most abundant (2340 GOs), followed by molecular function (MF, 1750 GOs) and cellular component (CC, 104 GOs) categories (Figure 2b). Within the BP category, metabolic processes (34.02%), cellular process (29.07%), and single-organism processes (16.28%) were most represented. Likewise, genes encoding binding proteins (33.40%) and genes encoding proteins related to catalytic activities (31.95%) were most abundant in the MF category. In the CC category, membrane (22.26%), membrane part (18.78%), cell (16.70%), and cell part (16.49%) were abundantly represented GO terms. In total, 10,960 (45% of the transcripts with NR blast hits for Viridiplantae) transcripts were assigned to different COG functional categories (Table S1). The largest group is represented by the serine theorine protein kinase (COG0515, 2014 hits, 18%), followed by leucine rich repeat (COG4886, 491 hits, 13%), ankyrin repeat (COG0666, 192 hits). A few other clusters, such us chromatin structures and dynamics, cell motility, extracellular structures are underrepresented or absent.

3.4. *Abundance of Transcripts Related to Terpene Functional Annotation*

A total of 186 transcripts were identified as being part of monoterpenoid biosynthetic processes (GO:0016099), triterpenoid biosynthetic processes (GO:0016104), including pentacyclic (GO:0019745) and tetracyclic triterpenoids (GO:0010686), terpenoid transport (GO:0046865), terpene synthase activity (GO:0010333) and terpene/terpenoid biosynthetic processes (GO:0046246 and GO:0016114, respectively). The lengths of these transcripts varied between 232 bp and 4278 bp, with a median contig length of 736 bp. In total, we found approximately 540.2 transcripts per million (TPM) expressed for all terpene and terpenoid functional annotation (Figure 3). Transcripts associated with the synthesis of monoterpenoids and terpenoid transport were relatively highly expressed with single transcripts having double the abundance relative to the mean TPM expressed in *Protium copal.*

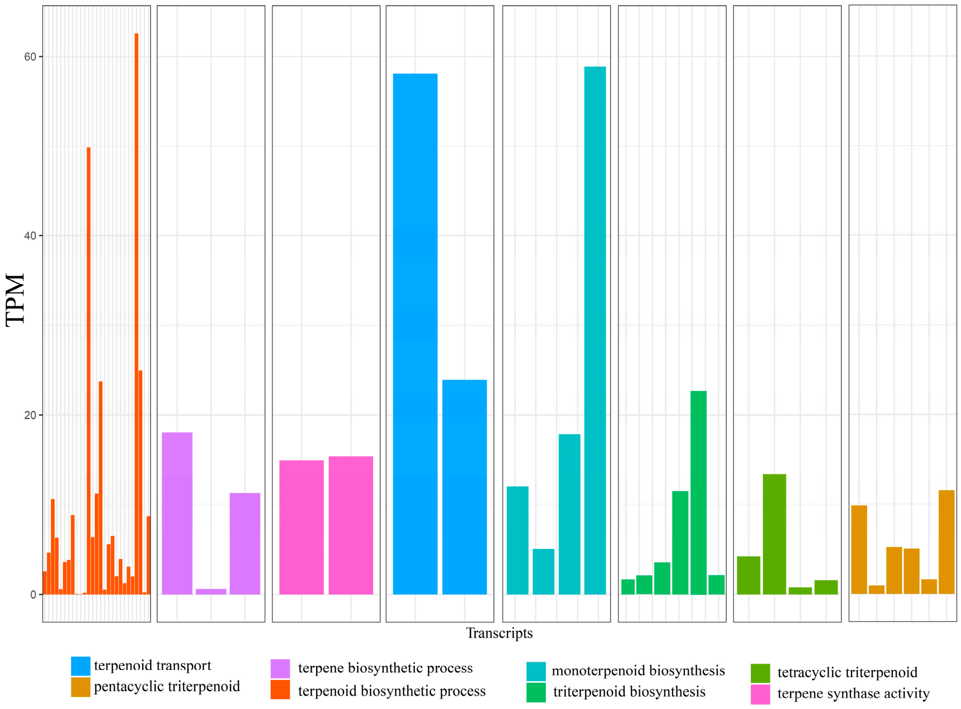

Figure 3. Expected abundance of transcripts expressed in transcripts per million (TPM) for different GO terms in *Protium copal* (Burseraceae). TPM is the abundance one would expect to find in a pool of a million transcripts.

3.5. *Phylogenetic Validation of Terpene Synthase Gene Annotations*

Fifty-seven *P. copal* transcripts were identified as putative TPS genes (Figure 4). With the exception of the outgroup specimens in TPSd all of the subfamilies were monophyletic. While support values are low within the TPSa and b subfamilies, the topology of the phylogeny is congruent with previous phylogenetic studies of the TPS gene family (see [4,7,30]). We conclude that the terpene genes identified in the *P. copal* transcriptome are consistent with the annotation of other model terpene-rich plant species (e.g., Citrus and Arabidopsis).

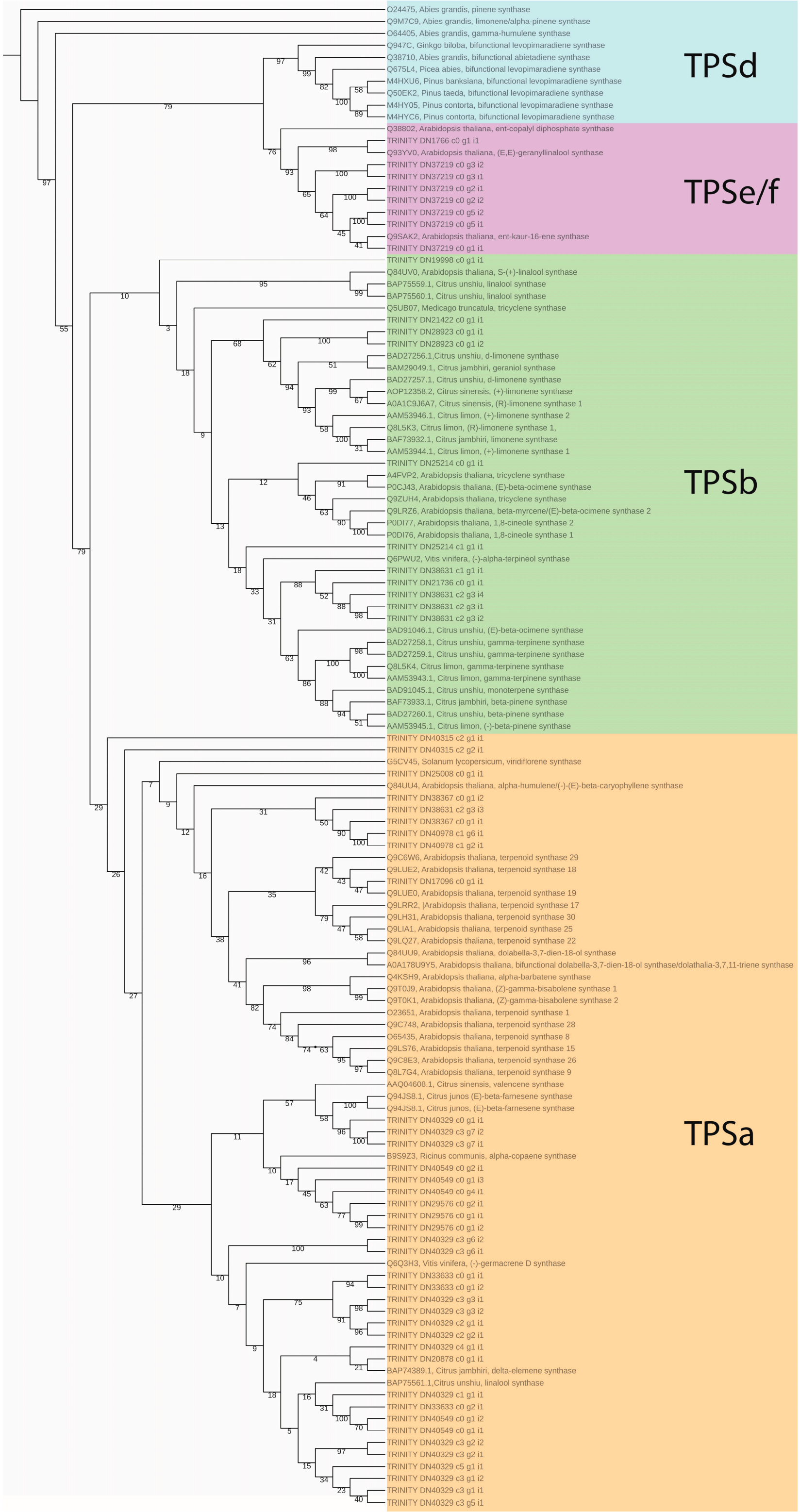
O24475, Abies grandis, pinene synthase
Q9M7C9, Abies grandis, limonene/alpha-pinene synthase
O64405, Abies grandis, gamma-humulene synthase
Q947C, Ginkgo biloba, bifunctional levopimaradiene synthase
Q38710, Abies grandis, bifunctional abietadiene synthase
Q675L4, Picea abies, bifunctional levopimaradiene synthase
M4HXU6, Pinus banksiana, bifunctional levopimaradiene synthase
Q50EK2, Pinus taeda, bifunctional levopimaradiene synthase
M4HY05, Pinus contorta, bifunctional levopimaradiene synthase
M4HYC6, Pinus contorta, bifunctional levopimaradiene synthase
TPSd
Q38802, Arabidopsis thaliana, ent-copalyl diphosphate synthase
TRINITY DN1766 c0 g1 i1
Q93YV0, Arabidopsis thaliana, (E,E)-geranyllinalool synthase
TRINITY DN37219 c0 g3 i2
TRINITY DN37219 c0 g3 i1
TRINITY DN37219 c0 g2 i1
TRINITY DN37219 c0 g2 i2
TRINITY DN37219 c0 g5 i2
TRINITY DN37219 c0 g5 i1
Q9SAK2, Arabidopsis thaliana, ent-kaur-16-ene synthase
TRINITY DN37219 c0 g1 i1
TPSe/f
TRINITY DN19998 c0 g1 i1
Q84UV0, Arabidopsis thaliana, S-(+)-linalool synthase
BAP75559.1, Citrus unshiu, linalool synthase
BAP75560.1, Citrus unshiu, linalool synthase
Q5UB07, Medicago truncatula, tricyclene synthase
TRINITY DN21422 c0 g1 i1
TRINITY DN28923 c0 g1 i1
TRINITY DN28923 c0 g1 i2
BAD27256.1,Citrus unshiu, d-limonene synthase
BAM29049.1, Citrus jambhiri, geraniol synthase
BAD27257.1, Citrus unshiu, d-limonene synthase
AOP12358.2, Citrus sinensis, (+)-limonene synthase
A0A1C9J6A7, Citrus sinensis, (R)-limonene synthase 1
AAM53946.1, Citrus limon, (+)-limonene synthase 2
Q8L5K3, Citrus limon, (R)-limonene synthase 1,
BAF73932.1, Citrus jambhiri, limonene synthase
AAM53944.1, Citrus limon, (+)-limonene synthase 1
TRINITY DN25214 c0 g1 i1
A4FVP2, Arabidopsis thaliana, tricyclene synthase
P0CJ43, Arabidopsis thaliana, (E)-beta-ocimene synthase
Q9ZUH4, Arabidopsis thaliana, tricyclene synthase
Q9LRZ6, Arabidopsis thaliana, beta-myrcene/(E)-beta-ocimene synthase 2
P0DI77, Arabidopsis thaliana, 1,8-cineole synthase 2
P0DI76, Arabidopsis thaliana, 1,8-cineole synthase 1
TRINITY DN25214 c1 g1 i1
Q6PWU2, Vitis vinifera, (-)-alpha-terpineol synthase
TRINITY DN38631 c1 g1 i1
TRINITY DN21736 c0 g1 i1
TRINITY DN38631 c2 g3 i4
TRINITY DN38631 c2 g3 i1
TRINITY DN38631 c2 g3 i2
BAD91046.1, Citrus unshiu, (E)-beta-ocimene synthase
BAD27258.1, Citrus unshiu, gamma-terpinene synthase
BAD27259.1, Citrus unshiu, gamma-terpinene synthase
Q8L5K4, Citrus limon, gamma-terpinene synthase
AAM53943.1, Citrus limon, gamma-terpinene synthase
BAD91045.1, Citrus unshiu, monoterpene synthase
BAF73933.1, Citrus jambhiri, beta-pinene synthase
BAD27260.1, Citrus unshiu, beta-pinene synthase
AAM53945.1, Citrus limon, (-)-beta-pinene synthase
TPSb
TRINITY DN40315 c2 g1 i1
TRINITY DN40315 c2 g2 i1
G5CV45, Solanum lycopersicum, viridiflorene synthase
TRINITY DN25008 c0 g1 i1
Q84UU4, Arabidopsis thaliana, alpha-humulene/(-)-(E)-beta-caryophyllene synthase
TRINITY DN38367 c0 g1 i2
TRINITY DN38631 c2 g3 i3
TRINITY DN38367 c0 g1 i1
TRINITY DN40978 c1 g6 i1
TRINITY DN40978 c1 g2 i1
Q9C6W6, Arabidopsis thaliana, terpenoid synthase 29
Q9LUE2, Arabidopsis thaliana, terpenoid synthase 18
TRINITY DN17096 c0 g1 i1
Q9LUE0, Arabidopsis thaliana, terpenoid synthase 19
Q9LRR2, |Arabidopsis thaliana, terpenoid synthase 17
Q9LH31, Arabidopsis thaliana, terpenoid synthase 30
Q9LIA1, Arabidopsis thaliana, terpenoid synthase 25
Q9LQ27, Arabidopsis thaliana, terpenoid synthase 22
Q84UU9, Arabidopsis thaliana, dolabella-3,7-dien-18-ol synthase
A0A178U9Y5, Arabidopsis thaliana, bifunctional dolabella-3,7-dien-18-ol synthase/dolathalia-3,7,11-triene synthase
Q4KSH9, Arabidopsis thaliana, alpha-barbatene synthase
Q9T0J9, Arabidopsis thaliana, (Z)-gamma-bisabolene synthase 1
Q9T0K1, Arabidopsis thaliana, (Z)-gamma-bisabolene synthase 2
O23651, Arabidopsis thaliana, terpenoid synthase 1
Q9C748, Arabidopsis thaliana, terpenoid synthase 28
O65435, Arabidopsis thaliana, terpenoid synthase 8
Q9LS76, Arabidopsis thaliana, terpenoid synthase 15
Q9C8E3, Arabidopsis thaliana, terpenoid synthase 26
Q8L7G4, Arabidopsis thaliana, terpenoid synthase 9
AAQ04608.1, Citrus sinensis, valencene synthase
Q94JS8.1, Citrus junos (E)-beta-farnesene synthase
Q94JS8.1, Citrus junos, (E)-beta-farnesene synthase
TRINITY DN40329 c0 g1 i1
TRINITY DN40329 c3 g7 i2
TRINITY DN40329 c3 g7 i1
B9S9Z3, Ricinus communis, alpha-copaene synthase
TRINITY DN40549 c0 g2 i1
TRINITY DN40549 c0 g1 i3
TRINITY DN40549 c0 g4 i1
TRINITY DN29576 c0 g2 i1
TRINITY DN29576 c0 g1 i1
TRINITY DN29576 c0 g1 i2
TRINITY DN40329 c3 g6 i2
TRINITY DN40329 c3 g6 i1
Q6Q3H3, Vitis vinifera, (-)-germacrene D synthase
TRINITY DN33633 c0 g1 i1
TRINITY DN33633 c0 g1 i2
TRINITY DN40329 c3 g3 i1
TRINITY DN40329 c3 g3 i2
TRINITY DN40329 c2 g1 i1
TRINITY DN40329 c2 g2 i1
TRINITY DN40329 c4 g1 i1
TRINITY DN20878 c0 g1 i1
BAP74389.1, Citrus jambhiri, delta-elemene synthase
BAP75561.1,Citrus unshiu, linalool synthase
TRINITY DN40329 c1 g1 i1
TRINITY DN33633 c0 g2 i1
TRINITY DN40549 c0 g1 i2
TRINITY DN40549 c0 g1 i1
TRINITY DN40329 c3 g2 i2
TRINITY DN40329 c3 g2 i1
TRINITY DN40329 c5 g1 i1
TRINITY DN40329 c3 g1 i2
TRINITY DN40329 c3 g1 i1
TRINITY DN40329 c3 g5 i1
TPSa

Figure 4. Maximum likelihood phylogeny of 57 putative TPS genes in *P. copal* and known TPS genes from angiosperm and gymnosperm taxa. Bootstrap support values greater than 50 are shown.

3.6. Relative Diversity of Orthologous Terpene Gene Clusters in P. copal

Comparison of inferred proteins among the three species in OrthoVenn identified 107 gene clusters, with 58 clusters being shared by at least two of the three species. Protium copal was found to share a relatively high number of terpene genes with both *B. simaruba* and *B. sacra* transcriptome assemblies (32 terpene genes were shared with B. sacra and 35 genes were shared with *B. simaruba*, Figure 5a). More unique terpene gene clusters were identified in *P. copal* and *B. simaruba* relative to B. sacra. In addition, *P. copal* shared more orthologous gene clusters with *B. simaruba* than *B. sacra*. Finally, *P. copal* exhibited the highest diversity of terpene orthologous gene clusters in comparison to the other two Burseraceae species (Figure 5b).

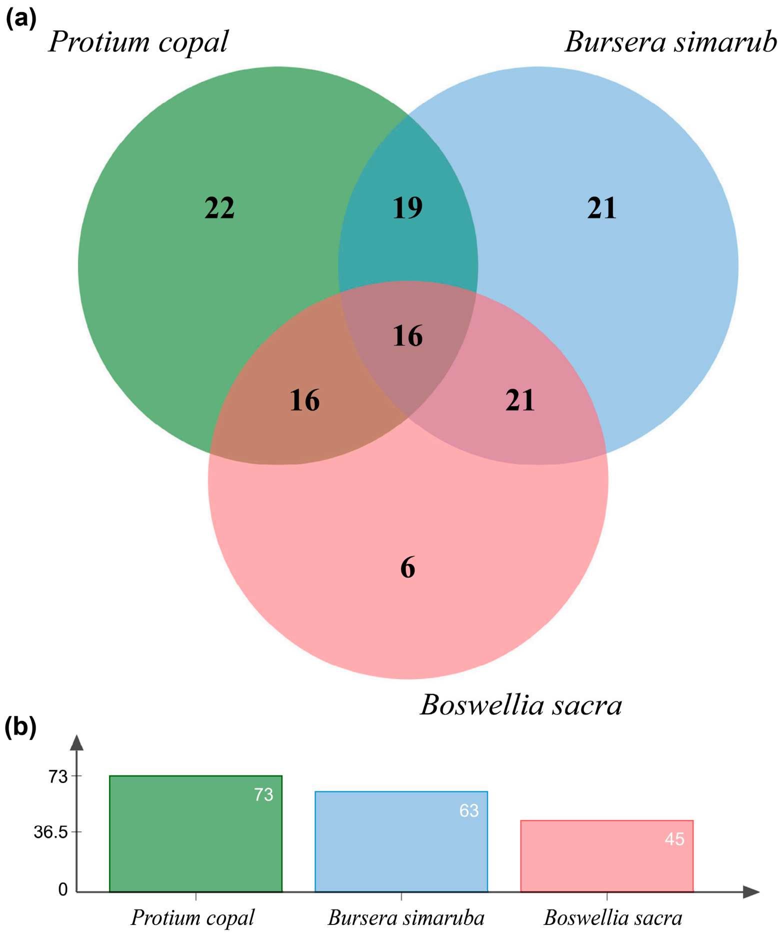

Figure 5. (**a**) Venn diagram of inferred terpene orthologous gene clusters for *Protium copal*, *Boswellia sacra*, and *Bursera simaruba* produced using OrthoVenn. (**b**) Total numbers of inferred terpene orthologous gene clusters for each species are given in the bar plot.

3.7. Simple Sequence Repeat (SSR) Marker Identification

A total of 11,480 repeat regions were identified across 9496 transcripts (15%). In addition, 1610 transcripts contained multiple SSRs (2.5%), and a total of 770 compound SSRs were identified in the *P. copal* transcriptome. Out of the 11,480 SSRs, 7308 mono-nucleotide (64%), 1936 di-nucleotide (17%), 2010 tri-nucleotide (18%), 112 tetra-nucleotide (1%), 53 penta-nucleotide (0.5%), and 61 hexa-nucleotide markers (0.5%) were identified. Seven SSRs were identified which were found within transcripts putatively involved in the terpenoid biosynthetic pathway and 3 of which were found in transcripts putatively involved in the isoprenoid biosynthetic pathway. Of these, five were mono-nucleotides, two were di-nucleotides, one was a tri-nucleotide, and two were compound SSR regions. The list of SSR markers identified are presented in Table S2.

3.8. SNP Discovery

A total of 64,510 SNPs was identified across 25,505 transcripts (Table S3). 36,893 SNPs were located in coding regions (CDS). 22,292 SNPs were classified as non-synonymous (35%), of which 22,159 are located in CDS regions. 74 SNPs were located across 30 transcripts involved in terpenoid biosynthesis pathway. Of these 74 SNPs, 28 were located in CDS regions and 14 SNPs were classified as non-synonymous, all of which were located in CDS regions. Fifty-seven were involved in isoprenoid biosynthesis genes, 16 in terpene synthase genes, and 1 in terpenoid biosynthesis genes.

4. Discussion

4.1. Novel Transcriptomic Resources for Burseraceae

Although Burseraceae is distributed throughout tropical and subtropical regions of the world, currently available transcriptomic information is limited to, *Boswellia* (frankincense) occurring in Asia and Africa, and *Bursera* (linaloe) occurring in Central America [31]. Here, we generate the first transcriptome annotation of *Protium*, a remarkable genus that is pantropical but harbors over 170 plant species in the Neotropics [32,33]. *Protium* is globally known for the aromatic and medicinal properties of resins and essential oils [9]. The availability of a comprehensive leaf transcriptome for *Protium copal* is the first step toward the development of genomics and medical applications. The transcriptomic data generated in this study provide useful resources to explore the functional aspects of Burseraceae resinous chemicals and investigate their genetic associations. Enzymes responsible for the synthesis of various secondary metabolites described in this species may be identified from the provided set of annotated gene domains, assembled transcripts or even based on the raw sequencing reads.

In addition, *Protium copal* transcripts could be used to generate novel sequencing markers applied to population genetics and comparative phylogenetics studies. High-throughput sequencing based on transcriptome capture are designed to enrich target genomic regions [34] and these techniques have been commonly used to generate well-resolved phylogenies for tropical plant groups [35,36]. *Protium* is considered a monophyletic genus and one of the most dominant plant genera in the Amazon region [28]. The development of target sequencing regions based on the annotated transcriptome of *Protium copal* could be directly used to improve the estimates of species limits and genetic diversity within *Protium*. Further transcriptomic analysis on multiple tissues and specimens, rather than a single reference, is still necessary to circumvent sampling bias and to ensure that the majority of genetic diversity within *Protium copal* is fully captured. Furthermore, de novo construction of a pan-transcriptome for *Protium* could help to unravel variation in gene regulation and expression and provide additional candidate genes for the study of select genotypes.

4.2. Assembly and Annotation Quality

We found that the assembly quality of *P. copal* transcriptome is equivalent or better than other transcriptomes of plant species in Burseraceae [31]. Over 50% of the transcripts and predicted proteins were successfully assigned to genes by BLASTx and BLASTp searches and more than 75% of transcripts returned a homologous BLAST hit with an e-value $< 1 \times 10^{-5}$, an indication of the high assembly quality for the *Protium copal* transcriptome. The exN50 is calculated similarly to N50 used for genome assemblies except that it is limited to the top most highly expressed transcripts that represent a 50% of the total normalized expression data. In contrast to whole-genome assemblies, transcriptomes might not achieve contigs with high exN50 values and the most highly expressed transcripts may not be the longest ones. However, exN50 for *Protium copal* is superior to other transcriptome assemblies in Burseraceae, and our statistics are comparable to recently published high-quality plant transcriptomes, such as [37–39]. Moreover, BUSCO represents a more appropriate measure to assess transcriptome quality by quantifying the presence of conserved orthologs in an assembly [15]. In comparison to other published transcriptomes in the Burseraceae (*Boswellia sacra* and *Bursera simaruba*), the *Protium copal* transcriptome had the least amount of missing single-copy orthologs, indicating a relatively complete assembly, and ~60% of orthologs were found in the assembly, indicating a relatively high-quality assembly. Finally, our phylogenetic analysis of putative TPS genes with known TPS genes from across angiosperms is congruent with previously published TPS phylogenies and further validates the annotations of these transcripts.

4.3. Identification of SSR Markers and SNPs in P. copal

SSR markers are frequently designed from transcriptomic assemblies providing a suitable source for genetic diversity assessment. Despite being derived from coding DNA regions, which tend to be evolutionarily conserved, SSRs developed from transcriptomes are considered a valuable genomic resource for studying the genetic structure of plant populations [40]. Here, we have made available a database of SSR markers expected to be polymorphic within *P. copal* that will be useful for genetic studies on this important medicinal plant. According to the Kissplice results, 22,293 transcripts were identified by having non-synonymous SNPs and 14,521 transcripts with synonymous SNPs. In addition, most of the SNPs (ca. 66%) are found in multiple assembled isoforms. This database presents a large collection of putative expressed genes that can be used in further genetic linkage and genome association analysis. In addition, primers can be easily designed for targeting specific genes in *Protium* and closely related genomes.

4.4. High Similarity with the Citrus Genome

The number of putative genes identified in the *P. copal* transcriptome was within the range of putative functional genes identified in other angiosperms [41–43]. As expected, the percent identity of assembled transcriptomes tends to increase as comparisons are made with plants at the same order, family or genus level [44]. Within our transcriptome assembly, over 38% of the transcripts were successfully annotated against one or more species of *Citrus*, a plant crop with important genomic resources. *Citrus* is a domesticated plant within the Rutaceae family [45], and like Burseraceae, is a member of the Sapindales order. The Citrus genome database is an open-source genome funded by the USDA and NSF agencies to enable basic and applied genomics, genetics, breeding and disease research [46]. Plant families in the Sapindales (e.g., Anacardiaceae, Burseraceae, Rutaceae) are known for producing a diverse suite of aromatic chemicals. Terpene gene families are found in species with specialized structures for storing volatile terpenes [47], such as *Citrus*, which accumulate volatiles in oil glands. Our transcriptome annotation includes several coding genes also found in *Citrus* that are useful for identifying proteins involved in the different steps of the terpene biosynthesis pathway.

4.5. Diversity of Terpene Biosynthetic Genes

Thousands of different terpenoid compounds are produced by plants through the expression of terpene synthase (TPS) genes. Terpenoids are characterized by an isoprenoid chemical structure and include derivatives with various functional groups [5]. The TPS gene family is classified according to phylogenetic relationships into eight subfamilies, which comprise mono-, sesqui-, di- and triterpene synthases [47]. In this study, the annotation of the *P. copal* transcriptome revealed terpenoid genes with different biosynthetic pathways (e.g., monoterpenoids, sesquiterpenes and triterpenoid) and genes responsible for terpene cell transportation. Terpenoid synthase activity expressed in *P. copal* is primarily responsible for the synthesis of linear terpenes (e.g., isopentenyl-PP, geranyl-PP, farnesyl-PP and geranylgeranyl-PP) containing varying numbers of isoprene units. Triterpenoid genes annotated in *P. copal* regulate the chemical reactions and pathways, resulting in six isoprene units and 4 or 5 carbon rings (tetracyclic and pentacyclic synthases). In addition, TPS genes found in *P. copal* results in the formation of monoterpenoids having a C10 molecule skeleton. We also found genes that are directly associated with the movement of terpenoids into or between conduit cells.

The origin and evolution of plant secondary metabolites in Burseraceae are an emerging theme in plant phytochemistry [8]. Although there are a few hundred terpenoids produced by almost all plants, the vast majority of terpenoids are restricted to a given lineage, or even a single species, and new terpenoids continue to be discovered in many plants [5]. The coevolutionary arms race hypothesis [48] has long been invoked to explain the diversity of terpene metabolites in aromatic plants like Burseraceae [49], in which specialized metabolites are predicted to diversify as a response of coevolution between plants and specialized herbivores. Defense against natural enemies is a well-recognized function of plant terpenoids and the production of terpenes represents a major evolutionary advantage against natural enemies [6]. Because plants in the tropics experience much higher levels of herbivory and pathogen attack compared to other biomes [10] it might be expected that when comparing terpene gene diversity among closely related taxa those found in tropical biomes would have a greater diversity of genes than those inhabiting other biomes. When we compared the transcriptomes of Burseraceae species inhabiting hyper diverse tropical regions (*P. copal* and *B. simaruba*), we found a greater amount of unique terpene orthologous genes as compared to *B. sacra*, an endemic taxon from arid regions in the North of Africa and Southern Middle East. This result was despite the fact that the *B. simaruba* transcriptome was significantly more incomplete than the *P. copal* and *B. sacra* transcriptomes. These results suggest that similar selection pressures in the shared ranges of *B. simaruba* and *P. copal* could possibly be responsible for their high level of terpene diversity relative *B. sacra*.

In addition to biotic pressures, the presence of a large number of genes already known to be involved in terpene biosynthesis could also underlie the ability to synthesize large numbers of terpenoids and increases the probability that new terpene genes will evolve. As a result of the complex chemical and physical properties of terpene metabolites, the diversity of terpenoids found across lineages may also be driven by other factors besides herbivore defense due to their role in primary metabolic functioning (e.g., electron transport chains depend on terpene association) and reproduction (e.g., chemical signaling to pollinators) [5]. Although examples of diverse terpenoid compounds have been found in some model organisms, the study of their evolution requires comparison with closely related species that occupy different ecological niches. Here, we provided a useful transcriptome annotation of terpene genes for *Protium* that will encourage future studies aiming to understand the evolution and diversification of terpene secondary metabolites in Burseraceae, one of the largest resinous plant families that produce a diverse array of terpene-related chemistry.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/5/392/s1, Table S1: Most recent Trinotate annotation report of the *Protium copal* (Burseraceae) transcriptome. Table S2: List of transcript-SSR markers along with their motifs and location in the *Protium copal* (Burseraceae) transcriptome. Table S3: List of transcripts of *Protium copal* (Burseraceae) containing putative SNPs and their respective Kissplice annotation.

Author Contributions: Conceptualization, G.D., V.S.S., T.M.M. and P.V.A.F; Methodology, G.D., T.M.M. and V.S.S.; Validation, G.D. and V.S.S.; Formal Analysis, G.D., T.M.M. and V.S.S.; Resources, T.M.M., D.C.D and P.V.A.F.; Data Curation, D.C.D; Writing—Original Draft Preparation, G.D., T.M.M. and V.S.S.; Writing—Review & Editing, T.M.M., D.C.D, P.V.A.F.; Visualization, G.D.; Supervision, P.V.A.F.; Project Administration, T.M.M.; Funding Acquisition, T.M.M. and P.V.A.F.

Funding: This research was financially supported by National Science Foundation Grants DEB-1254214 (P.V.A.F.) and DEB-1311117 (T.M.M and P.V.A.F.) and the first author was funded by the Brazilian program Science without Borders CNPq 202437/2011-6.

Acknowledgments: We thank Chodon Sass, Ke Bi and Lydia Smith for their careful assistance during laboratory work and bioinformatic analysis. We are thankful for the logistical support given by the curatorial staff in the New York Botanical Garden and the University and Jepson Herbaria. Also, we would like to express gratitude to the National Council for Scientific and Technological Development in Brazil and the National Science Foundation in the United States for funding this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Langenheim, J.H. *Plant Resins: Chemistry, Evolution, Ecology, and Ethnobotany*; No. 620.1924 L275p; Timber Press: Portland, OR, USA, 2003.
2. Stacey, R.J.; Cartwright, C.R.; McEwan, C. Chemical characterization of ancient mesoamerican copal resins: Preliminary results. *Archaeometry* **2006**, *48*, 323–340. [CrossRef]
3. Cheng, A.X.; Lou, Y.G.; Mao, Y.B.; Lu, S.; Wang, L.J.; Chen, X.Y. Plant terpenoids: Biosynthesis and ecological functions. *J. Integr. Plant Biol.* **2007**, *49*, 179–186. [CrossRef]
4. Chen, F.; Tholl, D.; Bohlmann, J.; Pichersky, E. 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.* **2011**, *66*, 212–229. [CrossRef]
5. Pichersky, E.; Raguso, R.A. Why do plants produce so many terpenoid compounds? *New Phytol.* **2018**, *220*, 692–702. [CrossRef] [PubMed]
6. Kessler, A.; Heil, M. The multiple faces of indirect defences and their agents of natural selection. *Funct. Ecol.* **2011**, *25*, 348–357. [CrossRef]
7. Zapata, F.; Fine, P.V.A. Diversification of the monoterpene synthase gene family (TPSb) in Protium, a highly diverse genus of tropical trees. *Mol. Phylogenetics Evol.* **2013**, *68*, 432–442. [CrossRef]
8. Salazar, D.; Lokvam, J.; Mesones, I.; Vásquez, M.; Ayarza, J.M.; Fine, P.V.A. Origin and maintenance of chemical diversity in a species-rich tropical tree lineage. *Nat. Ecol. Evol.* **2018**, *2*, 983–990. [CrossRef]
9. Rüdiger, A.L.; Siani, A.C.; Junior, V.V. The chemistry and pharmacology of the South America genus Protium Burm. f. (Burseraceae). *Pharmacogn. Rev.* **2007**, *1*, 93–104.
10. Coley, P.D.; Barone, J.A. Herbivory and plant defenses in tropical forests. *Annu. Rev. Ecol. Syst.* **1996**, *27*, 305–335. [CrossRef]
11. Daly, D.C. A Taxonomic Revision of 'Protium"('Burseraceae') in Eastern Amazonia and the Guianas. Ph.D. Dissertation, University Microfilms, Ann Arbor, MI, USA, 1987.
12. Bolger, A.M.; Marc, L.; Bjoern, U. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
13. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
14. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525. [CrossRef]
15. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
16. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494. [CrossRef] [PubMed]
17. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, 29–37. [CrossRef]
18. Powell, S.; Szklarczyk, D.; Trachana, K.; Roth, A.; Kuhn, M.; Muller, J.; Arnold, R.; Rattei, T.; Letunic, I.; Doerks, T.; et al. EggNOG v3. 0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **2011**, *40*, 284–289. [CrossRef] [PubMed]
19. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef]
20. Götz, S.; García-Gómez, J.M.; Terol, J.; Williams, T.D.; Nagaraj, S.H.; Nueda, M.J.; Robles, M.; Talón, M.; Dopazo, J.; Conesa, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **2008**, *36*, 3420–3435. [CrossRef]
21. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
22. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **2017**. [CrossRef]

23. Kuraku, S.; Zmasek, C.M.; Nishimura, O.; Katoh, K. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. *Nucleic Acids Res.* **2013**, *41*, 22–28. [CrossRef]
24. Miller, M.A.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of the 2010 Gateway Computing Environments Workshop (GCE), New Orleans, LA, USA, 14 November 2010; pp. 1–8.
25. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **2006**, *23*, 127–128. [CrossRef] [PubMed]
26. Wang, Y.; Coleman-Derr, D.; Chen, G.; Gu, Y.Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **2015**, *43*, 78–84. [CrossRef] [PubMed]
27. Thiel, T.; Michalek, W.; Varshney, R.; Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor. Appl. Genet.* **2003**, *106*, 411–422. [CrossRef]
28. Sacomoto, G.A.; Kielbassa, J.; Chikhi, R.; Uricaru, R.; Antoniou, P.; Sagot, M.F.; Lacroix, V. K is splice: De-novo calling alternative splicing events from RNA-seq data. *BMC Bioinform.* **2012**, *13*, 5.
29. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef]
30. Bohlmann, J.; Meyer-Gauen, G.; Croteau, R. Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 4126–4133. [CrossRef]
31. Matasci, N.; Hung, L.H.; Yan, Z.; Carpenter, E.J.; Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Ayyampalayam, S.; Barker, M.; et al. Data access for the 1,000 Plants (1KP) project. *Gigascience* **2014**, *3*, 17. [CrossRef]
32. Daly, D.C.; Fine, P.V.A.; Martínez-Habibe, M.C. Burseraceae: A model for studying the Amazon flora. *Rodriguésia* **2012**, *63*, 21–30. [CrossRef]
33. Fine, P.V.A.; Zapata, F.; Daly, D.C. Investigating processes of neotropical rain forest tree diversification by examining the evolution and historical biogeography of the Protieae (Burseraceae). *Evolution* **2014**, *68*, 1988–2004. [CrossRef]
34. Bi, K.; Vanderpool, D.; Singhal, S.; Linderoth, T.; Moritz, C.; Good, J.M. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genom.* **2012**, *13*, 403. [CrossRef]
35. Prata, E.M.; Sass, C.; Rodrigues, D.P.; Domingos, F.M.; Specht, C.D.; Damasco, G.; Ribas, C.C.; Fine, P.V.A.; Vicentini, A. Towards integrative taxonomy in Neotropical botany: Disentangling the Pagamea guianensis species complex (Rubiaceae). *Bot. J. Linn. Soc.* **2018**, *188*, 213–231. [CrossRef]
36. Sass, C.; Iles, W.J.; Barrett, C.F.; Smith, S.Y.; Specht, C.D. Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* **2016**, *4*, e1584. [CrossRef]
37. He, R.; Kim, M.J.; Nelson, W.; Balbuena, T.S.; Kim, R.; Kramer, R.; Crow, J.A.; May, G.D.; Thelen, J.J.; Soderlund, C.A.; et al. Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, Phragmites australis (Poaceae), reveals genes involved in invasiveness and rhizome specificity. *Am. J. Bot.* **2012**, *99*, 232–247. [CrossRef] [PubMed]
38. Tian, X.J.; Long, Y.; Wang, J.; Zhang, J.W.; Wang, Y.Y.; Li, W.M.; Peng, Y.F.; Yuan, Q.H.; Pei, X.W. De novo transcriptome assembly of common wild rice (Oryza rufipogon Griff.) and discovery of drought-response genes in root tissue based on transcriptomic data. *PLoS ONE* **2015**, *10*, e0131455. [CrossRef] [PubMed]
39. Li, M.; Liang, Z.; Zeng, Y.; Jing, Y.; Wu, K.; Liang, J.; He, S.; Wang, G.; Mo, Z.; Tan, F.; et al. De novo analysis of transcriptome reveals genes associated with leaf abscission in sugarcane (Saccharum officinarum L.). *Bmc Genom.* **2016**, *17*, 195. [CrossRef]
40. Taheri, S.; Lee Abdullah, T.; Yusop, M.; Hanafi, M.; Sahebi, M.; Azizi, P.; Shamshiri, R. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. *Molecules* **2018**, *23*, 399. [CrossRef]
41. Falara, V.; Akhtar, T.A.; Nguyen, T.; Spyropolou, E.A.; Bleeker, P.M.; Schauvinhold, I.; Matsuba, Y.; Bonini, M.E.; Schilmiller, A.L.; Last, R.L.; et al. The tomato terpene synthase gene family. *Plant Physiol.* **2011**, *157*, 770–789. [CrossRef] [PubMed]

42. Nieuwenhuizen, N.J.; Green, S.A.; Chen, X.; Bailleul, E.J.D.; Matich, A.J.; Wang, M.Y.; Atkinson, R.G. Functional genomics reveals that a compact terpene synthase gene family can account for terpene volatile production in apple. *Plant Physiol.* **2013**, *161*, 787–804. [CrossRef] [PubMed]
43. Külheim, C.; Padovan, A.; Hefer, C.; Krause, S.T.; Köllner, T.G.; Myburg, A.A.; Degenhardt, J.; Foley, W.J. The Eucalyptus terpene synthase gene family. *BMC Genom.* **2015**, *16*, 450–466. [CrossRef]
44. Geniza, M.; Jaiswal, P. Tools for building de novo transcriptome assembly. *Curr. Plant Biol.* **2017**, *11*, 41–45. [CrossRef]
45. Wu, G.A.; Terol, J.; Ibanez, V.; López-García, A.; Pérez-Román, E.; Borredá, C.; Domingo, C.; Tadeo, F.R.; Carbonell-Caballero, J.; Alonso, R.; et al. Genomics of the origin and evolution of Citrus. *Nature* **2018**, *554*, 311. [CrossRef]
46. Xu, Q.; Chen, L.L.; Ruan, X.; Chen, D.; Zhu, A.; Chen, C.; Bertrand, D.; Jiao, W.B.; Hao, B.H.; Lyon, M.P.; et al. The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.* **2013**, *45*, 59. [CrossRef] [PubMed]
47. Nagegowda, D.A. Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation. *Febs Lett.* **2010**, *584*, 2965–2973. [CrossRef]
48. Ehrlich, P.R.; Raven, P.H. Butterflies and plants: A study in coevolution. *Evolution* **1964**, *18*, 586–608. [CrossRef]
49. Becerra, J.X.; Noge, K.; Venable, D.L. Macroevolutionary chemical escalation in an ancient plant–herbivore arms race. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18062–18066. [CrossRef]