agronomy

MDPI

Article

Identification of Terpene-Related Biosynthetic Gene Clusters in Tobacco through Computational-Based Genomic, Transcriptomic, and Metabolic Analyses

Roel C. Rabara [1], Chengalrayan Kudithipudi [2] and Michael P. Timko [1,*]

[1] Department of Biology, University of Virginia, Charlottesville, VA 22903, USA; roel.rabara.contractor@altria.com
[2] Altria Client Services, Center for Research and Technology, Richmond, VA 23219, USA; chengalrayan.kudithipudi@altria.com
* Correspondence: mpt9g@virginia.edu

Abstract: Terpenes and terpenoids contribute aroma and flavor that influence consumer preferences in selecting plant-based products. Computational identification of biosynthetic gene clusters (BGCs) in plants can pave the way for future biosynthetic genetic engineering. Using integrative genomic, transcriptomic, and metabolic pathway annotation analyses, 35 BGCs were identified in tobacco with high confidence. Among the 35 BGCs identified, 7 were classified as terpene biosynthesis-related BGCs. Two BGCs found on C13 and C14 chromosomes belonged to terpene and saccharide-terpene biosynthetic classes that were only 93 Mb and 189 Kb apart, respectively. Other clusters have lengths ranging from 120 Kb (Cluster 9) to 1.6 Mb (Cluster 18). Each cluster contained five (Cluster 21) to twenty genes (Cluster 32), and the number of terpene synthase genes present in the clusters also varied from one (Clusters 18 and 21) to eight (Cluster 32). Gene expression profiling using diurnal and topping transcriptome datasets identified co-expressing genes within modules and varying levels of expression among modules as represented by the normalized enrichment score measured in each module. The positions pinpointed from these computational analyses will allow for the more efficient modifications of specific genes and BGCs for the development of tobacco-based products with improved aroma and flavor.

Keywords: biosynthetic gene clustering; specialized metabolites; terpene synthase; co-expressing genes; specialized metabolic pathways; tobacco

Citation: Rabara, R.C.; Kudithipudi, C.; Timko, M.P. Identification of Terpene-Related Biosynthetic Gene Clusters in Tobacco through Computational-Based Genomic, Transcriptomic, and Metabolic Analyses. *Agronomy* **2023**, *13*, 1632. https://doi.org/10.3390/agronomy13061632

Academic Editor: Chengsheng Zhang

Received: 23 May 2023
Revised: 14 June 2023
Accepted: 16 June 2023
Published: 18 June 2023

1. Introduction

Recent advances in the plant genomics era herald bright prospects for biosynthetic genetic engineering. Researchers now have better molecular tools that can facilitate further understanding of a plant at a whole system level. Researchers can further explore the innate evolutionary capacity of plants to produce a vast array of primary and secondary metabolites (SMs) that consequentially develop the multiple complex metabolic pathways involved in their biosynthesis. Some plant metabolites have figured prominently in agrochemical, pharmaceutical, and nutraceutical product development [1,2]. Plant metabolites have been estimated to range between 200,000 and 1 million [3], awaiting the further discovery of their characteristics and potential utilities in agriculture, medicine, and other industries.

Perhaps hampering the possibilities of these biotechnological interventions is the lack of understanding of plant biosynthesis toward improving crop yields and creating synthetic processes for cost-efficient production [1]. Suggestions have been made that linking genomic and transcriptomic attributes to metabolic output could help control the formation of SMs [3,4]. Thus, the key to modifying plant metabolism for crop improvement and biotechnological innovations is a profound understanding of gene organization and expression [5].

Genetic and biochemical studies have demonstrated that plant secondary metabolism can often include the physical clustering of genes involved in the production of SMs [1]. Additionally, general metabolism involves pathways conserved across most plant species, and the production of SMs tends to be more species- and genera-specific, regardless of evidence of the inter family sharing of pathways [6]. Gene clustering occurs in a distinct chemical pathway adjacent to at least three non-homologous genes in a genome's biosynthetic pathway [1], where these gene-encoding enzymes interact to define and modify a scaffold toward catalyzing the pathway end of product formation [7].

The existence of physical gene clustering in specialized metabolic pathways is well known in bacteria and fungi [1,8]. Before the genomic era, documentation of biosynthetic gene clusters (BGC) in plants was more limited, and data showing gene co-expressions in secondary metabolic pathways were often clearly less evident. However, as high-quality genomes of important crops and wild grasses become available, more BGCs are being discovered [9]. Recently, the co-expression of genes in BGCs leading to the biosynthesis of specialized molecules involved in plant defense and virulence has been well characterized [10].

Perhaps the first-ever study of plant gene physical clustering in a specialized metabolite pathway was in maize (*Zea mays*) that revealed the synthesis of the allelochemical 2,4-dihydroxy-1,4-benzoxazin-3-one (DIMBOA) [11]. Aside from maize, BGCs forming DIMBOA and its 7-methoxy analog DIMBOA have now been identified in several plants like barnyard grass (*Echinochloa crus-galli*) and wheat (*Triticum aestivum*) [12].

In recent years, the clustering of non-homologous and co-localized metabolic genes for secondary metabolite biosynthesis has yielded multiple examples of gene clusters for SMs [9,13,14]. These include compounds such as momilactone A and phytocassanes A-E in *Oryza sativa* [15], avenacin A-1 in *Avena strigosa* [16], the triterpenes thalianol and marneral in *Arabidopsis thaliana* [17], the alkaloid noscapine in *Papaver somniferum*, a-chaconine/a-solanine in *Solanum tuberosum*, the cyanogenic glucosides lotaustralin and linamarin in *L. japonicus*, and a-tomatine in *S. lycopersicum* [1].

Terpenes and terpenoid compounds are SMs that are nearly ubiquitous in plants where they play an essential role in the plants' central cellular processes, including photosynthesis, cell wall formation, electron transport, signaling, and membrane fluidity [18–20]. Additionally, terpenoids function in various ways in interactions between plants and their biotic and abiotic stressors in the environment [1,21]. The Dictionary of Natural Products Database records about 70,000 isolated terpenoid compounds [1,20]. The term "terpenoid" includes all molecules obtained from the condensation of the C5 precursor isopentenyl pyrophosphate (C5) and its allylic isomer dimethylallyl pyrophosphate (C5) [22,23]. Terpenes are biosynthesized through key enzymes called terpene synthases (TPSs) [24]. Historically and at present, there is significant interest in the regulation of *TPS* genes and their encoded enzymes and products owing to the importance of terpenes in commercial products such as rubber, taxol, artemisinin, labdanoid sclareol, carnosic acid, and carnosol, among others [19,22].

To date, terpenoid-related biosynthetic pathways are the predominant pathways where evidence of gene clustering has been found to be involved. For example, terpenoids in both *L. japonicus* [25] and *S. lycopersicum* [26] and diterpenoids, such as phytocassane, oryzalexin, and momilactone, in rice (*O. sativa*) [25], all appear to require some level of gene clustering. Similarly, sesquiterpenoids such as capsidiol in both *Capsicum* and *Nicotiana* species, rishitin in the *Solanum* species, and zealexin in maize (*Z. mays*) have been implicated in involving clustering of biosynthetic genes [27]. The identification of secondary metabolic gene clusters requires the use of various approaches, such as map-based cloning, genome mining, and forward and reverse genetics, as well as functional analysis to confirm new clusters [28] and bioinformatic tools like anti-SMASH 2.0 for genome mining [29].

For example, a gene cluster identified in *L. japonicus* was expressed in *N. benthamiana*, and its expression pattern was observed under various environmental and developmental conditions. Suggestively, the *L. japonicus* gene cluster functions in legume triterpene biosynthesis, with a possible role in plant development [25]. In another Solanum species study of the gene cluster for terpene biosynthesis, combined phylogenetic, genomic, and

biochemical analyses revealed divergent biochemical evolution and gene accreditation processes linked with metabolic diversification [26]. Moreover, gene clustering in plants facilitates the delineation of various plant biosynthetic pathways [10], including pathways for terpenes, that are the focus of this computational study.

Among the most widely used model plant species, *Nicotiana* species like tobacco (*Nicotiana tabacum*) are known to produce thousands of SMs, especially terpenoid compounds that serve important roles in plant growth and development as well as various plant biotic and environmental interactions [1,24].

While there are a limited number of examples of BGCs existing in tobacco involved in secondary metabolite formation, no previous attempt has been made to systematically examine BGCs in this species due to the complexity of the tobacco genome. The recently improved tobacco genome assembly [30] now makes it possible to comprehensively investigate the association of BGCs and secondary metabolism in tobacco. BGCs that will be identified in tobacco can be applied to other related important solanaceous crops like tomato by searching for clustered orthologs or syntenic regions [10], showing the importance of this study. The knowledge of BGCs can be used to facilitate pathway discovery, and elucidate existing biosynthetic pathways and the production of natural products through metabolic engineering [10].

In this study, we integrated genomic, transcriptomic, and metabolic pathway annotation analyses to identify BGCs in *N. tabacum* species with high confidence. Our main objective was to identify terpene-related BGCs in order to provide a deeper understanding of the characteristics of these gene clusters. We examined both the gene content of the BGCs and their correlated gene expression data to better understand the coordinated function of the BGCs during various plant growth and development stages. Our research findings could facilitate more efficient validations of terpene-related genes and BGCs toward specific gene alterations in tobacco for better terpene-related product development, such as aromatic and flavorful tobacco products, and for robust terpene-led plant defense mechanisms for biotic and abiotic stresses for improvement in crop production.

2. Materials and Methods

2.1. Terpene Synthase (TPS) Gene Family Investigation and Phylogenetic Analysis

The genome and the annotation files used in this study were downloaded from Sol Genomics Network (SGN, https://solgenomics.net/; accessed on 20 November 2020) [31], including genome sequences of *N. tabacum* cultivar 'K326' ver. 4.5 [30]. Members of *TPS* gene families (Pfam PF01397, PF03936) were extracted for subsequent sequence and phylogenetic analyses in these studies. The published TPS protein sequences from tomato (*Solanum lycopersicum*) [32] were used to anchor tobacco TPS for phylogenetic and molecular evolutionary analyses using MEGA ver. X [33]. Sequence alignment was performed using the built-in aligner CLUSTALW, and the phylogenetic tree was inferred by the Maximum Likelihood method using Jones–Taylor–Thornton (JTT) substitution model. Bootstrap method with 1000 replicates was used to test for the constructed phylogenetic tree. Motif analysis of TPS amino acid sequences was performed for each identified group using MEME SUITE v.5.4.1 following the default settings [34]. To link molecular functions to the TPS genes identified in the tobacco genome, a search for gene orthologs was performed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) reference database [35] using the amino acid sequence of the identified TPS.

2.2. RNA-Seq Data and Co-Expression Analysis

In this study, two RNA-Seq datasets were employed for gene expression analysis of TPS genes that were downloaded from the NCBI Gene Expression Omnibus (GEO) database. These are the RNA-Seq dataset on diurnal expression of genes in tobacco (GEO Accession: GSE95717 [30]) and the expression profile of axillary bud outgrowth after topping (GEO Accession: GSE153483 [36]).

The GSE95717 dataset comprises data generated from eight-week-old tobacco plants (cv. 'K326') grown under long day photoperiod (18/6 h) from which three tissues (root,

whole shoot, and shoot apex tissues) were harvested at Zeitgeber time ZT0, 6, 12 and 18 h (h) after light imposition to represent diurnal gene expression. The GSE153483 dataset is the transcriptomics data of tobacco axillary shoot outgrowth after 1 d, 3 d, and 5 d after topping. Sequence reads were mapped to publicly available reference tobacco genome using TopHat v2.0.12 [37]. Normalization was performed based on the length of the gene and the count of mapped reads to the gene.

After normalization, the RNA-Seq datasets were analyzed using the CEMiTool R package [38] following the default settings (p-value = 0.1 and Pearson correlation method) to identify genes that are co-expressing in the tobacco transcriptome. The CEMiTool allows for the visualization of individual gene expression across samples from different groups defined by the user and performs Gene Set Enrichment Analyses (GSEA) that showcase the module activity of each group of samples. The CEMiTool also provides for run-over representation analysis that pinpoints the top ten most connected genes (hubs) and defines module functions to create gene networks.

2.3. Identification of Biosynthetic Gene Clusters in Tobacco Genome

The complete assembled and annotated tobacco genome sequence ver. 4.5 [30] was downloaded from Sol Genomics Network database (https://solgenomics.net/; accessed on 20 November 2020). It was used as input sequence in plantiSMASH (http://plantismash.secondarymetabolites.org/; accessed on 20 November 2020) [39] to identify putative BGCs in the tobacco genome following the default settings. JBrowse 2 [40] was used to manually measure the location and distance of genes and clusters in K326 tobacco genome.

3. Results

3.1. Genome-Wide Identification, Phylogenetic Tree Construction, and Classification of Terpene Synthase in Tobacco Genome

Exactly 160 terpene synthases (TPSs) were identified (Figure 1) after analyzing the ~69,500 annotated protein sequences of tobacco [30]. Based upon their sequences and using the published tomato TPS [32] as an anchor, the tobacco *TPS* genes were assigned to one of five groups/clades (designated TPS-a, -b, -c, -e/f, and -g). Our findings concurred with known groupings of the *TPS* gene family in plants [41]. TPS-d and TPS-h were not present, wherein TPS-d is a gymnosperm-specific clade, and TPS-h is thus far specific to the lycopod *Selaginella moellendorffii* which indicates the lineage-specific expansion of the *TPS* family [32,41]. A total of 95 TPS genes belong to TPS-a, while 32, 11, 13, and 1 TPS genes clustered in the clades TPS-b, -c, -e/f, and -g, respectively (Supplementary Figure S1).

Results show all groups have the conserved DDxxD motif (see Supplementary Figure S1), except for TPS-c, which is characterized by a highly conserved DxDD motif (Supplementary Figure S1C). On the other hand, TPS-b has the typical DDxxD motif and the RRX_8W motif (Supplementary Figure S1B). TPS-d was not observed in tobacco which is consistent with published findings that TPS-d is only observed in gymnosperms [32,41].

To link the identified 160 TPS genes to molecular functions and established pathways, we searched our genes for ortholog groups in the KEGG database. Supplementary Figure S2 shows the summary of ortholog groups identified in the KEGG database. Out of 160 TPSs, 66 genes (46%) have annotations in the database, of which 64 genes are involved in the metabolism of terpenoids and polyketides, and 2 are involved in lipid metabolism.

3.2. Gene Expression Profile of Terpene Synthase Genes

Figure 2 shows the results of gene expression analyses of *TPS* genes from two tobacco datasets: diurnal expression of genes (GEO Acc. GSE95717) and expression profile of axillary bud outgrowth after topping (GEO Acc. GSE153483). Diurnal gene expression was both influenced spatially (type of tissue samples) and temporally (sampling time points), as reflected in Figure 3A. Notably, some *TPS* genes were highly expressed in the roots compared to those in the shoots and shoot apices (Figure 2A). Three *TPS* genes were observed showing expression only in the roots: *Nitab4.5_0000013g0430*, *Nitab4.5_0004597g0030*, and

Nitab4.5_0012291g0010. The cluster analysis of the gene expression profile revealed tissue-dependent expression as exhibited by the grouping of each tissue sample (Figure 2A).

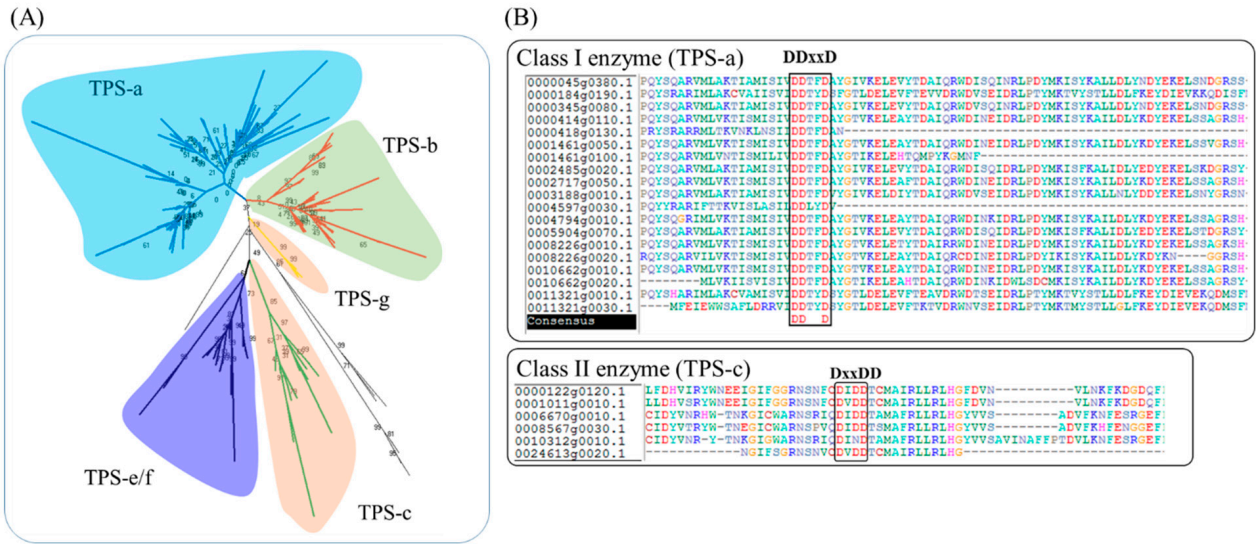

Figure 1. Phylogenetic analysis of terpene synthase (TPS) family in tobacco. (**A**) Unrooted phylogenetic tree of *TPS* gene family in tobacco genome with tomato TPS used as an anchor. (**B**) Conserved motif of class I (e.g., TPS-a) and class II (TPS-c) enzymes.

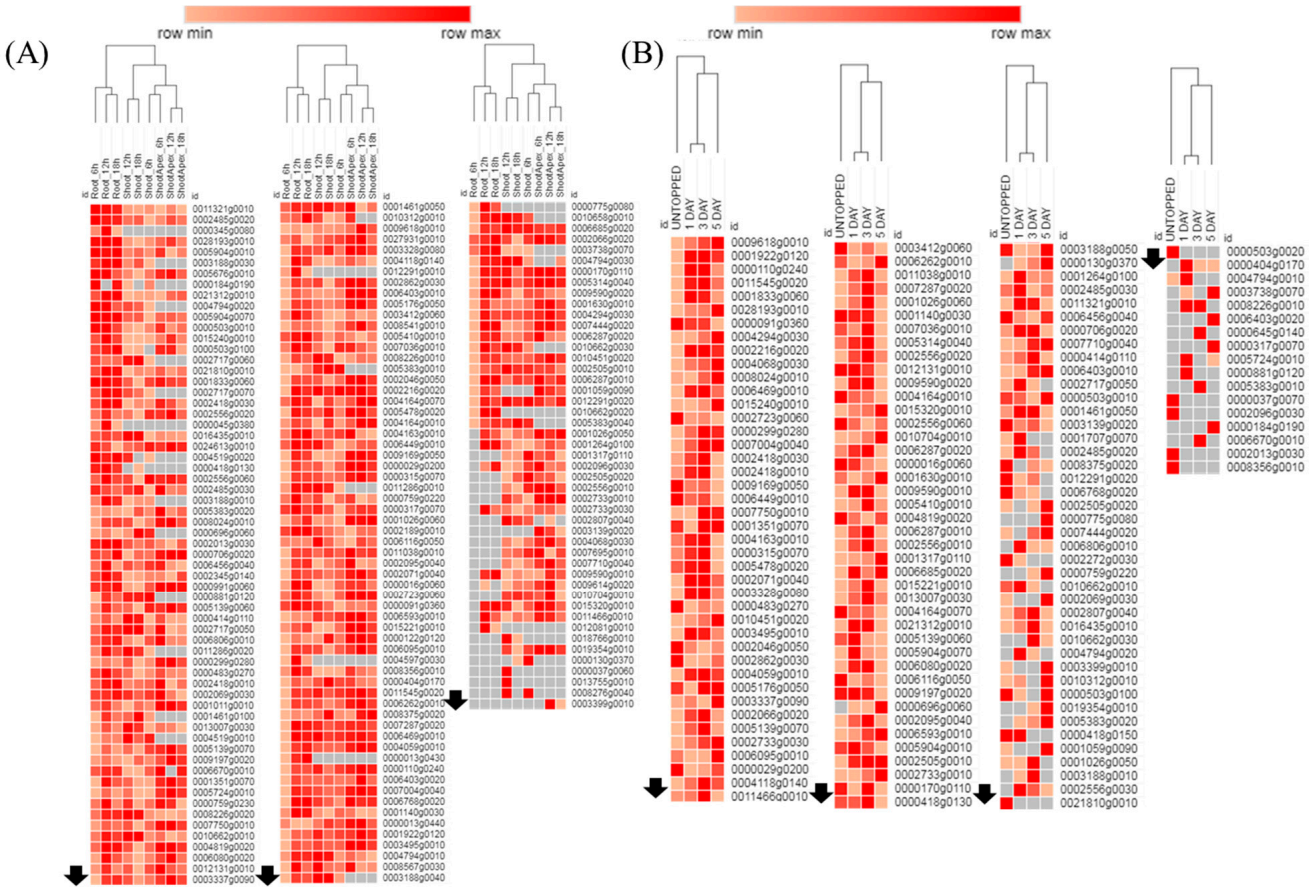

Figure 2. Gene expression profile of terpene synthase (*TPS*) genes in tobacco cv. 'K326'. (**A**) Diurnal expression of *TPS* genes in root, shoot, and shoot apex and (**B**) expression profile of *TPS* genes in axillary buds after topping.

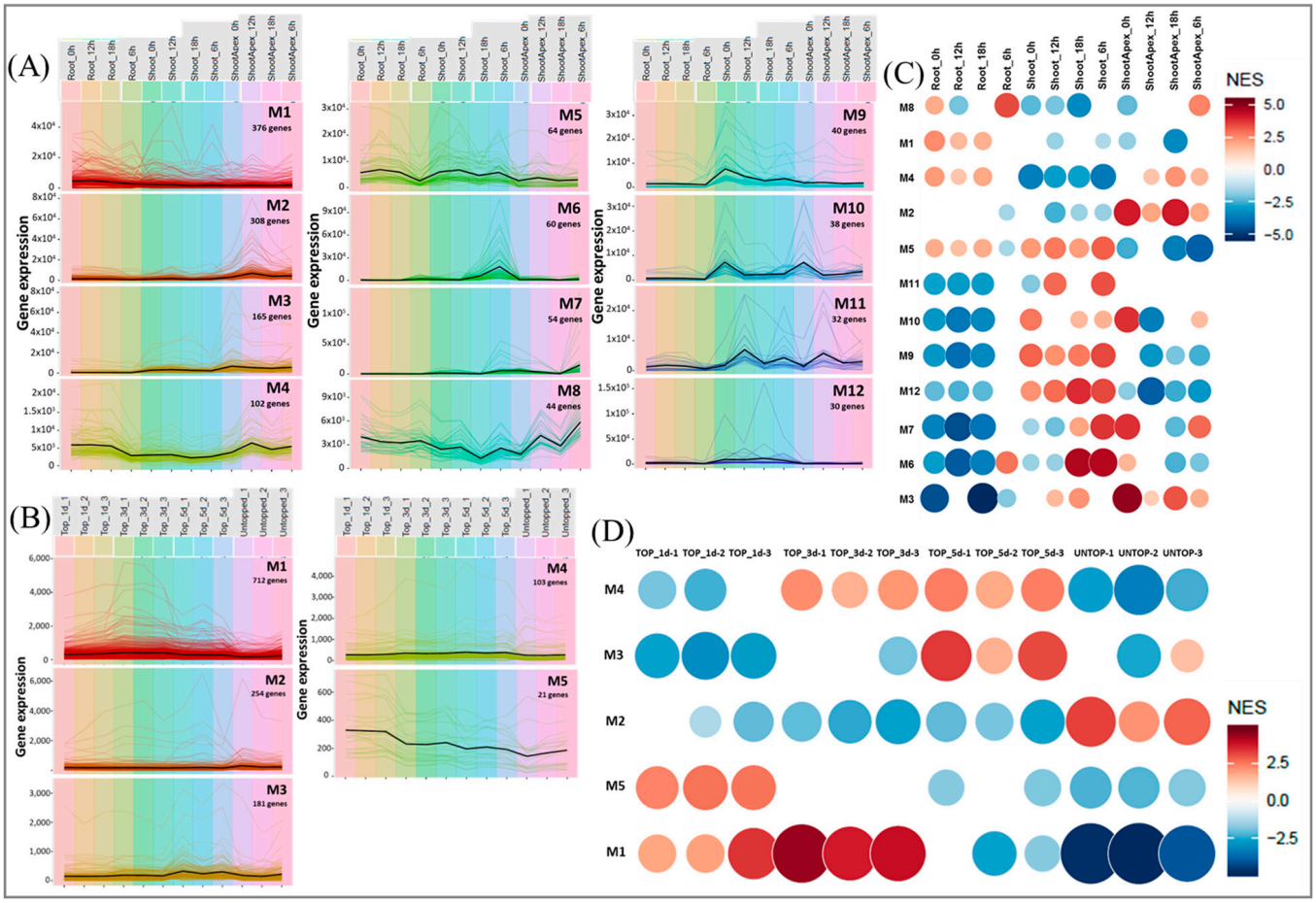

Figure 3. Gene expression modules in diurnal expression of tobacco genes (**A**,**B**) and in tobacco axillary buds after topping of tobacco plants (**C**,**D**). The expression levels of individual genes from each module are shown as colored lines. The black line represents the mean expression of all genes inside the module. Samples are shown on the *x*-axis and colored by classes (**A**,**C**). Gene Set Enrichment Analyses (GSEA) show the module activity (indicated by size of circle) on each class of sample, where color shows the normalized enrichment score (NES). Enrichment score (ES) reflects the degree to which a gene set is overrepresented on a ranked list of genes (**B**,**D**).

When the expression patterns of *TPS* genes were examined in the axillary bud after topping, they were found to be expressed in newly developing buds as early as one day after topping (Figure 2B). Six *TPS* genes, *Nitab4.5_0021810g0010*, *Nitab4.5_0008396g0010*, *Nitab4.5_0002013g0030*, *Nitab4.5_0000037g0070*, *Nitab4.5_0002096g0030*, and *Nitab4.5_0000503g0020* were not detected after topping. Cluster analysis showed that expression patterns of *TPSs* in axillary buds can be categorized into early (one and three days after topping) and late responses (five days after topping).

3.3. Identification of Co-Expression Modules in Tobacco Transcriptome Datasets

Using CeMiTool, genes that interact together in a global transcriptome data were analyzed and identified from two publicly available datasets (GEO Accessions: GSE95717 and GSE153483) downloaded from the NCBI GEO database.

After examining each individual gene expression in the two datasets, 12 co-expression modules were identified in the diurnal expression dataset (Figure 3A). Modules are defined as gene sets that have similar expression patterns, *viz.*, genes that are over-represented by specific pathways or changed in a specific sample group [38]. The number of gene members in each module ranged between 30 and 376 genes. Gene expression profiles of most of the modules were not affected by tissue samples and sampling time except for modules M8-M11. Module M8 showed that gene expression profiles in the roots were not affected

by time but were clearly affected in the shoot and the shoot apices. Similar trends were observed in M10 and M11, while in M9, fluctuation of gene expression due to temporal effect (time course sampling effect) was only observed in shoot samples. The expression profile of M9 genes in shoot tissue showed decreased expression at 6, 12, and 18 h, which may indicate that genes in this module were affected by the diurnal rhythm in tobacco plants. Biological rhythms, such as circadian rhythms, define the daily phase of biological processes, which include the organization of transcriptome to ensure coordinated responses at the appropriate time of the day at the cellular level [42].

Member genes in each identified module should show co-expression profiles. So, to further understand the level of the co-expression behavior of each module, we did a Pearson correlation coefficient (PCC) analysis. PCC is one of the most popular methods in co-expression analysis, in which it measures the tendency of expression levels between a pair of genes [43]. PCC scores range between −1 and +1, with negative scores meaning the two genes respond in the opposite manner, while positive scores mean both genes respond in the same manner [43]. The results indicate that the expression pattern of most genes in each module was positively correlated (Supplementary Figure S3A), validating their inclusion in their respective modules. Among the 12 modules, only M1 showed a significant number of negative PCC values that comprise 33% of the total values calculated. This indicates that a third of the genes exhibited an inverse correlation with each other. The gene expression profile for each module can be further explained by the gene activity in the gene set enrichment analysis (GSEA) shown in Figure 2B.

GSEA revealed that co-expressed modules were enriched in shoot and shoot apex samples across time points. A closer look at the module activity of M8 (Figure 3B) revealed a slight increase in gene expression in the root sample at ZT6h caused an increase in module activity and positive normalized enrichment score (NES), while a decrease in gene expression in the shoot sample at ZT18h resulted in an increase in module activity and negative NES values. Gene sets in M3 were highly overrepresented in shoot apex at the ZT0h time point, while the M6 gene sets were overrepresented at the ZT6 and ZT18 h time points.

In the case of expression profiles of axillary bud outgrowth after topping, only five modules were identified (Figure 3C). The number of genes co-expressed in each module ranged between 21 (M5) and 712 (M1). M1, being the largest module in terms of the number of genes in the module, is followed by M2 (254 genes), M3 (181 genes), M4 (103 genes), and M5 (21 genes). The expression profile of each module showed no significant fluctuation across sampling time points except in M5. Members of the M5 module showed an increase in gene expression after topping, then started to decline at three and five days after topping. Further analysis by PCC revealed the expression behavior of the five co-expression modules (Supplementary Figure S3B). M1 showed that 91% of calculated PCC values were positive, indicating that most members of this module exhibited a direct correlation of expression. Meanwhile, M2, M3, and M4 showed that 30% of pair-wise gene expression analyses were inversely correlated. In M5, only *nucleoside diphosphate kinase* (*NDPK*, *Nitab4.5_0003596g0010*) and an *aquaporin-like* gene (*AQP, Nitab4.5_0010813g0020*) showed an inverse correlation in their gene expression. *NDPKs* are ubiquitous enzymes involved in the synthesis and maintenance of nucleotide triphosphate (NTPs) pools [44]. *AQPs* are globally present across the plant kingdom and are members of the major intrinsic proteins involved in the selective transport of substrates needed for various biological processes [45].

GSEA analysis of this dataset revealed an increasing trend in module activities in M1, the biggest module identified (Figure 3D). In untopped plants, gene activities in the M1 module exhibited low activity but immediately started to increase a day after topping and peaked at 3 d, and then started to decline 5 days later. The decline in gene activities could be indicative that the plant has acclimated to having been topped. The M2 module, on the other hand, showed high activity before the plants were topped, but started to decline immediately after topping and further decreased its activities until after 5 d post-topping.

The M3 and M4 modules have similar activity profiles wherein in untopped plants, gene activities were initially low and increased 5 and 3 days after topping in M3 and M4 modules, respectively. The M5 module was somewhat different, having increased gene activities a day after topping and then rapidly declining in the succeeding days after topping.

3.4. Identification of Biosynthetic Gene Clusters in Tobacco Genome

The whole genome assembly of tobacco cv. 'K326' was used as input data for the identification of BGCs in tobacco using the plantiSMASH analysis platform. Table 1 shows the 35 BGCs identified through the analysis representing the seven biosynthetic product classes. Saccharide was the most abundant class (43%), followed by terpene (14%), saccharide-terpene (6%), polyketide (6%), alkaloid (6%), lignan (3%), with 23% categorized as an unclassified biosynthetic class (putative).

Figure 4 depicts the seven terpene-related BGCs identified from the tobacco genome. The seven terpene BGCs were identified in five chromosomes (chromosomes 7, 10, 13, 14, and 22). Both Chromosomes 13 and 14 have two BGCs, each belonging to terpene and saccharide-terpene biosynthetic classes. Cluster size ranges from 120 Kb (Cluster 9) to 1.6 Mb (Cluster 18), and the two clusters in Chromosomes 13 and 14 are only 93 Mb and 189 Kb apart, respectively. The number of gene members for each cluster ranges from 5 (Cluster 21) to 20 members (Cluster 32). The number of terpene synthase genes present in the clusters also varied from one (Cluster 18 and 21) to eight (Cluster 32) (Figure 4). One important gene family in terpene biosynthesis is *CYP* (Cytochrome P450). *TPSs* and *CYPs* are considered the primary drivers of terpene diversification [46]. *CYP* is one of the largest gene families in plants and is widely involved in various biosynthesis of plant natural products [47]. It is also a key driver in alkaloid diversification due to the formation and re-arrangement of alkaloid scaffolds that are catalyzed by P450s [48]. Notably, only clusters 21, 22, 24, and 32 among the terpene-related BGCs contain the *CYP* gene (Table 1). Cluster 21 consists of one pair of *TPS* genes (*Nitab4.5_0000404g0170.1*) and *CYP* (*Nitab4.5_0000404g0200.1*), which are ~110 kb apart. Cluster 22, which is the second terpene cluster on Chromosome 13, contains seven *TPSs* and two *CYPs*. The first *CYP* is ~113 kb from the closest *TPS*, while the second *CYP* is ~5 kb from its closest *TPS*. In cluster 24, there are two pairs of *TPSs* (*Nitab4.5_0000037g0100.1* and *Nitab4.5_0000037g0110.1*) and *CYP* (*Nitab4.5_0000037g0200.1* and *Nitab4.5_0000037g0210.1*). The two *TPSs* are 58 kb apart from each other, while the two *CYP* genes are farther apart from each other at 287 kb. In this cluster, the *TPS* and *CYP* are ~361 kb apart. On the other hand, Cluster 32, which has the highest number of gene members (20 genes), includes four *CYPs* and eight *TPS*. The closest observed pairing of *TPS/CYP* was ~41 kb, while the farthest distance was the ~181 kb region.

To further understand the gene expression profile of members of each cluster, we conducted a PCC analysis, and the data are presented in Supplementary Figure S4. Three of the seven BGCs identified (Cluster 21, 22, and 23) have positive PCC values ranging between 70 and 73%. On the other hand, the other four clusters have positive PCC values ranging between 45 and 57%. This indicates that most of the gene members in Cluster 21, 22, and 23 have a direct linear relationship.

Table 1. Biosynthetic gene clusters (BGC) identified from the tobacco 'K326' genome, including core domains for each cluster, using plantiSMASH analysis platform. Terpene-related BGCs are in bold letters.

Cluster	Chromosome	Biosynthetic Class	Core Domains Identified
1	Nt01	Saccharide	Transferase, Glycosyltransferase Family 28 N-Terminal Domain
2	Nt02	Saccharide	2OG-Fe(II) Oxygenase, Cellulose Synthase, glycosyltransferase family 28 n-terminal domain, and Cytochrome P450
3	Nt03	Saccharide	CO Esterase, Glycosyltransferase Family 28 N-Terminal Domain
4	Nt03	Alkaloid	AMP Binding, Str Synthase
5	Nt03	Putative	2OG-Fe(II) Oxygenase, DIOX N
6	Nt04	Putative	Methyl Transferase 2, Cytochrome P450
7	Nt05	Saccharide	Transferase, Glycosyltransferase Family 28 N-Terminal Domain
8	Nt06	Lignan	Dirigent, Cytochrome P450
9	**Nt07**	**Terpene**	Amino Oxidase, Terpene Synthase, and Terpene Synthase C
10	Nt07	Polyketide	Chal Sti Synthase C, Chal Sti Synthase N, Methyl Transferase 7, and SE
11	Nt07	Putative	Amino Oxidase, Omt
12	Nt08	Putative	DIOX N, FA Desaturase2, Prenyl Transferase, and Cytochrome P450
13	Nt08	Saccharide	Amino Oxidase, Amino Trans3, DIOX N, Transferase, Glycosyltransferase Family 28 N-Terminal Domain, and Cytochrome P450
14	Nt09	Saccharide	Glycosyltransferase Family 28 N-Terminal Domain, Cytochrome P450
15	Nt09	Saccharide	Epimerase, Transferase, and Glycosyltransferase Family 28 N-Terminal Domain
16	Nt09	Saccharide	Lipoxygenase, Glycosyltransferase Family 28 N-Terminal Domain, and Cytochrome P450
17	Nt10	Saccharide	Glycosyltransferase Family 28 N-Terminal Domain, Cytochrome P450
18	**Nt10**	**Terpene**	Terpene Synthase, Terpene Synthase C, and Transferase
19	Nt10	Saccharide	2OG-Fe(II) Oxygenase, Amino Transferase, Cellulose Synthase, DIOX N, Glycosyltransferase Family 28 N-Terminal Domain, and Cytochrome P450
20	Nt12	Saccharide	Lyase Aromatic, Glycosyltransferase Family 28 N-Terminal Domain, and Polyprenyl Synthase
21	**Nt13**	**Saccharide-Terpene**	Glycosyl Transferase1, SQ Hop Cyclase C, and Cytochrome P450
22	**Nt13**	**Terpene**	AMP-Binding, Terpene Synthase, Terpene Synthase C, and Cytochrome P450
23	**Nt14**	**Terpene**	AMP-Binding, Terpene Synthase, and Terpene Synthase C
24	**Nt14**	**Saccharide-Terpene**	Terpene Synthase, Terpene Synthase C, Glycosyltransferase Family 28 N-Terminal Domain, and Cytochrome P450
25	Nt14	Putative	2OG-Fe(II) Oxygenase, DIOX-N, and Cytochrome P450
26	Nt14	Putative	2OG-Fe(II) Oxygenase, AMP-Binding, Amino Oxidase, and SQS-PSY
27	Nt17	Putative	2OG-Fe(II) Oxygenase, DIOX-N
28	Nt18	Saccharide	Amino Trans3, Glycosyltransferase Family 28 N-Terminal Domain
29	Nt19	Putative	Acetyl Transferase I, Peptidase10
30	Nt20	Alkaloid	Cu Amine Oxidase, Cytochrome P450
31	Nt20	Polyketide	Chalcone And Stilbene Synthases, Methyl Transferase 2, and Cytochrome P450
32	**Nt22**	**Terpene**	Prenyl Transferase, Terpene Synthase, Terpene Synthase C, and Cytochrome P450
33	Nt24	Saccharide	Epimerase, Glycosyltransferase Family 28 N-Terminal Domain
34	Nt24	Saccharide	Peptidase S10, Glycosyltransferase Family 28 N-Terminal Domain
35	Nt24	Saccharide	Transferase, Glycosyltransferase Family 28 N-Terminal Domain, and Cytochrome P450

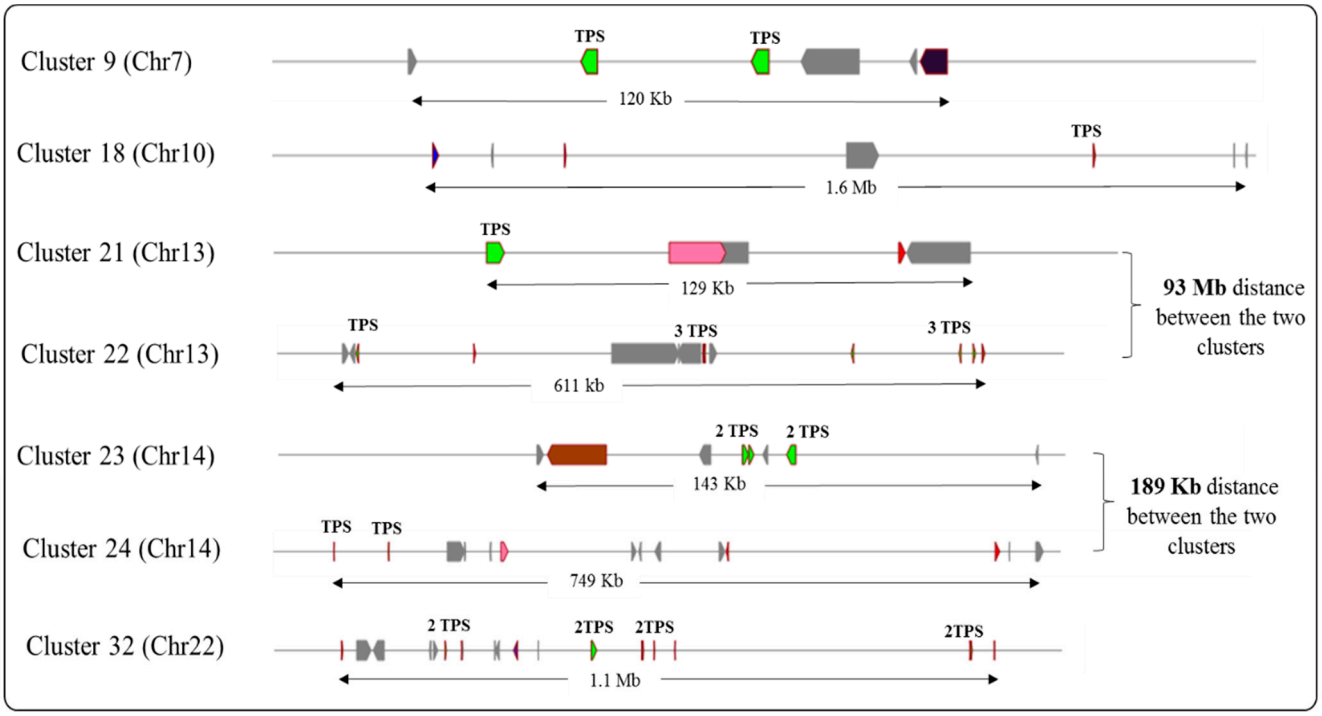

Figure 4. Terpene-related biosynthetic gene clusters (BGC) identified in *N. tabacum* cv. 'K326' genome. The key BGCs identified, along with the *TPS* gene family members present, and the length of the whole cluster are shown. Numbers in parentheses correspond to the chromosome numbers listed in Table 1. Light green color represents TPS gene, gray color represents BGC-related genes, and other remaining colors represent core biosynthetic genes.

4. Discussion

In this study, we identified 160 *TPS* genes through sequence analyses of the whole tobacco genome. Generally, *TPSs* are classified into two broad classes (I and II) based on their chemical reaction and their generated products. Class I utilizes a trinuclear metal cluster to trigger the ionization of an isoprenoid diphosphate substrate to yield an allylic cation and inorganic pyrophosphate. Class II terpenoid cyclase relies on a general acid (an aspartic acid side chain) to protonate the terminal carbon−carbon double bond of an isoprenoid substrate to yield a tertiary carbocation [41,49]. Full-length TPS sequence is usually characterized by conserved domains at N-terminal (Pfam ID PF01397) and C-terminal (Pfam ID PF03936) ends. The N-terminal domain contains the RRX_8W motif, and the C-terminal domain has the DDxxD and NSE/DTE motifs [24].

Our phylogenetic analysis resulted in five clades (a, b. c, e/f, and g), with clades d and h not observed because these clades are specific to gymnosperm and *S. moellendorffii*, respectively [41]. Except for clade c, all clades showed the DDxxD motif. Members of clade c have the DxDD motif that is important for the protonation-initiated cyclization reaction mechanism in class II TPSs [50]. Class II TPSs are represented by class II terpene synthases that synthesize ent-copalyl diphosphate and copal-8-ol diphosphate.

The use of genome mining tools utilizing hidden Markov models to identify genomic loci to predict multiple enzymes associated with SMs [51] has been reported [39]. Using the plantiSMASH analysis platform, a genomic survey of putative BGCs was performed on tobacco cv. 'K326' genome. The platform algorithm identified gene clusters by first scanning all genes predicted to encode biosynthetic enzymes, including specific signature protein domains that belonged to scaffold-generating enzymes specific for a class of biosynthetic pathways on their comprehensive profile hidden Markov Models (pHMMs) [39]. Then, the algorithm looked for the co-occurrence of at least three biosynthetic enzyme-coding

genes belonging to at least two different enzyme types and subsequently included any flanking genes.

The genomic scan of the tobacco genome yielded 35 BGCs belonging to seven biosynthetic classes, of which 20% (seven clusters) belong to a terpene-related biosynthetic class. These terpene BGCs varied in length ranging from 120 Kb to 1.6 Mb. This varying length of clusters identified by the plantiSMASH platform may be attributed to its detection strategy of using dynamic cut-off rather than the static kilobase cut-off usually used in gene cluster detection in fungal genomes [39]. Opting for dynamic cut-off rather than static distance cut-off addresses the inherent characteristics of high variability in gene densities in plant genomes. The use of static distance cut-off in the detection of gene clusters in plants has been employed before by Mao et al. [52], wherein their group detected 43 terpene gene clusters in 35 plant species after scanning 107 plant genomes within a 100 kb window. Variability in the core domains in each terpene cluster varied, with only four (clusters 21, 22, 24, and 32) out of seven terpene clusters containing genes belonging to the *CYP* family. Chen et al. [53] noted that *TPS* and *CYP* always locate and function as clusters, indicating the importance of *CYP* in a terpene cluster. Clustering of *TPS/CYP* was shown to be optimal in the 50 kb region after scanning 17 representative monocot and dicot genomes in the 30–200 kb regions [46].

The biosynthesis of terpenoids chiefly engages TPSs and another enzyme class called cytochrome P450 (CYP) [53]. Terpenoid biosynthesis involves three major steps: (1) conversion of linear isoprenyl diphosphates to linear or cyclic terpene olefins or terpene alcohols by TPSs, resulting in the diversity of terpene backbone structures observed in nature; (2) oxidation of the terpenes by CYP oxygenases as well as dioxygenases and various other types of dehydrogenases; and (3) further functionalization by the addition of a variety of substituents, including acyl-, glycosyl-, benzoyl-, or even alkaloid-groups [54].

The pairing of *TPS-CYP* is prevalent in multiple plant genomes [46], where they always locate and function as metabolic gene clusters [26,46,55]. For example, different mechanisms of pathway assembly in eudicots and monocots have been observed in the matching and mixing of individual *TPS* and *CYP* genes [46]. These genes jointly form the core components of terpene biosynthetic pathways, and their interaction generates a vast array of diverse terpene structures [41,46]. CYPs' involvement in the dealkylation, oxidation, dehydration, C-C cleavage, decarboxylation, desaturation, dimerization, isomerization, reduction, and ring extension reactions made them one of the key contributors to chemical diversity in plant metabolites [54].

To examine whether the genes present in the BGCs predicted by bioinformatic approaches are coregulated, we analyzed the diurnal and topping transcriptome datasets to identify modules in BGCs with co-expressing modules. Our analysis of spatial and temporal gene expression in tobacco identified co-expressing modules with varying expression profiles. Activities of each module were also observed as represented by the NES measured in each module. However, a survey of genes in each module demonstrated that none of the genes from the predicted BGCs formed a co-expression module. However, when we analyzed the expression profile of each predicted BGC using PCC analysis, some of the genes from each BGC were found to be co-expressed (Supplemental Figure S4). This observation has also been previously reported when Wisecaver et al. [56] analyzed published bioinformatically predicted BGCs and suggested that most of the predicted plant BGCs are not genuine SM pathways. However, some SMs are products of plant response to the environment, and the diurnal and topping gene expression datasets probably did not reflect the co-expression patterns that we expected from the predicted BGCs. Validation through metabolic profiling could support the predicted BGCs and further explain the results of the co-expression analysis.

5. Conclusions

Using integrative genomic, transcriptomic, and metabolic pathway annotation analyses, we identified, with high confidence, 35 BGCs in *N. tabacum* (tobacco) species. We

examined both the gene content of the BGC and correlated gene expression data to better understand the coordinated function of the BGCs in select plant growth and development stages under varied developmental programs. Notably, two BGCs belonging to terpene and saccharide-terpene biosynthetic classes found in C13 and C14 chromosomes were only 93 Mb and 189 kb apart, respectively. Only these two chromosomes have two BGCs, which may be an indicator of the importance of these chromosomes in terpene biosynthesis. Overall, our research findings provide knowledge that could facilitate more efficient and targeted modifications of terpene-related genes and BGCs in attempts to direct alterations in tobacco for the improvement in aroma in tobacco-based product development and plant defense mechanisms against biotic and abiotic stresses. Overall, computational analysis of BGCs provides a guide for researchers to identify and validate candidate genes that can be used in enhancing biosynthetic pathways for targeted products.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/agronomy13061632/s1. Supplementary Figure S1. Identified consensus motif in each member of the TPS clades (A) TPS-a, (B) TPS-b, (C) TPS-c, (D) TPSe/f, and (E) TPS-g. Supplementary Figure S2. KEGG annotation of the 160 TPS identified in the tobacco genome. Functional classification of the annotated TPSs (A). KEGG orthology (KO) annotation of the 66 TPS genes. Supplementary Figure S3. Heatmap of Pearson correlation coefficient (PCC) analysis of gene expression of members of each identified co-expression module in (A) diurnal global transcriptome and (B) topping transcriptome of tobacco genes. Supplementary Figure S4. Pearson correlation coefficient (PCC) analysis of gene expression of members of each terpene-related biosynthetic gene cluster in diurnal global transcriptome of tobacco genes. Cluster number represents the cluster number indicated in Supplementary Table S1. Supplementary Table S1. Gene members of each terpene-related biosynthetic gene cluster (BGC) in tobacco genome.

Author Contributions: Conceptualization and design of experiments, M.P.T. and R.C.R.; performed the research, R.C.R.; assisted with the bioinformatic analysis, C.K.; data analysis; R.C.R., C.K. and M.P.T.; writing—original draft, review and editing, R.C.R., C.K. and M.P.T.; supervision, M.P.T.; project administration, M.P.T.; funding acquisition, M.P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by sponsored research project GI15099 awarded to MPT. The funding sponsors (Altria Client Services) had no role in the design of the study, collection, analyses, or interpretation of data, writing of the manuscript; or in the decision to publish the results.

Data Availability Statement: All data belonging to this manuscript are available from the corresponding author upon request. Public data used in the analyses are available on the NCBI GEO website (https://www.ncbi.nlm.nih.gov/geo/).

Acknowledgments: The authors thank Hai Liu for his assistance with gene expression analysis and Tatyana Kotova for technical assistance.

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

References

1. Nutzmann, H.W.; Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **2014**, *26*, 91–99. [CrossRef] [PubMed]
2. Scherlach, K.; Hertweck, C. Mining and unearthing hidden biosynthetic potential. *Nat. Commun.* **2021**, *12*, 3864. [CrossRef] [PubMed]
3. Rabara, R.C.; Tripathi, P.; Rushton, P.J. Comparative Metabolome Profile between Tobacco and Soybean Grown under Water-Stressed Conditions. *BioMed Res. Int.* **2017**, *2017*, 3065251. [CrossRef] [PubMed]
4. Fiehn, O. Metabolomics—The link between genotypes and phenotypes. In *Functional Genomics*; Town, C., Ed.; Springer: Dordrecht, The Netherlands, 2002; pp. 155–171.
5. Delli-Ponti, R.; Shivhare, D.; Mutwil, M. Using Gene Expression to Study Specialized Metabolism—A Practical Guide. *Front. Plant Sci.* **2021**, *11*, 625035. [CrossRef]
6. Moghe, G.D.; Kruse, L.H. The study of plant specialized metabolism: Challenges and prospects in the genomics era. *Am. J. Bot.* **2018**, *105*, 959–962. [CrossRef]
7. Osbourn, A. Gene clusters for secondary metabolic pathways: An emerging theme in plant biology. *Plant Physiol.* **2010**, *154*, 531–535. [CrossRef]

8. Chu, H.Y.; Wegel, E.; Osbourn, A. From hormones to secondary metabolism: The emergence of metabolic gene clusters in plants. *Plant J.* **2011**, *66*, 66–79. [CrossRef]
9. Guo, J.; Ren, Y.; Tang, Z.; Shi, W.; Zhou, M. Characterization and expression profiling of the ICE-CBF-COR genes in wheat. *PeerJ* **2019**, *7*, e8190. [CrossRef]
10. Polturak, G.; Osbourn, A. The emerging role of biosynthetic gene clusters in plant defense and plant interactions. *PLoS Pathog.* **2021**, *17*, e1009698. [CrossRef] [PubMed]
11. Frey, M.; Chomet, P.; Glawischnig, E.; Stettner, C.; Grun, S.; Winklmair, A.; Eisenreich, W.; Bacher, A.; Meeley, R.B.; Briggs, S.P.; et al. Analysis of a chemical plant defense mechanism in grasses. *Science* **1997**, *277*, 696–699. [CrossRef] [PubMed]
12. Wu, D.; Jiang, B.; Ye, C.Y.; Timko, M.P.; Fan, L. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. *Plant Commun.* **2022**, *3*, 100320. [CrossRef]
13. Boycheva, S.; Daviet, L.; Wolfender, J.L.; Fitzpatrick, T.B. The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **2014**, *19*, 447–459. [CrossRef]
14. Nützmann, H.-W.; Huang, A.; Osbourn, A. Plant metabolic clusters—From genetics to genomics. *New Phytol.* **2016**, *211*, 771–789. [CrossRef]
15. Wilderman, P.R.; Xu, M.; Jin, Y.; Coates, R.M.; Peters, R.J. Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiol.* **2004**, *135*, 2098–2105. [CrossRef]
16. Qi, X.; Bakht, S.; Leggett, M.; Maxwell, C.; Melton, R.; Osbourn, A. A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8233–8238. [CrossRef]
17. Field, B.; Fiston-Lavier, A.-S.; Kemen, A.; Geisler, K.; Quesneville, H.; Osbourn, A.E. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16116–16121. [CrossRef] [PubMed]
18. Gershenzon, J.; Dudareva, N. The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **2007**, *3*, 408–414. [CrossRef] [PubMed]
19. Pichersky, E.; Raguso, R.A. Why do plants produce so many terpenoid compounds? *New Phytol.* **2018**, *220*, 692–702. [CrossRef] [PubMed]
20. Vavitsas, K.; Fabris, M.; Vickers, C.E. Terpenoid Metabolic Engineering in Photosynthetic Microorganisms. *Genes* **2018**, *9*, 520. [CrossRef]
21. Rabara, R.C.; Tripathi, P.; Reese, R.N.; Rushton, D.L.; Alexander, D.; Timko, M.P.; Shen, Q.J.; Rushton, P.J. Tobacco drought stress responses reveal new targets for Solanaceae crop improvement. *BMC Genom.* **2015**, *16*, 484. [CrossRef]
22. Bruckner, K.; Tissier, A. High-level diterpene production by transient expression in Nicotiana benthamiana. *Plant Methods* **2013**, *9*, 46. [CrossRef] [PubMed]
23. Heiling, S.; Schuman, M.C.; Schoettner, M.; Mukerjee, P.; Berger, B.; Schneider, B.; Jassbi, A.R.; Baldwin, I.T. Jasmonate and ppHsystemin regulate key Malonylation steps in the biosynthesis of 17-Hydroxygeranyllinalool Diterpene Glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. *Plant Cell* **2010**, *22*, 273–292. [CrossRef]
24. Jiang, S.-Y.; Jin, J.; Sarojam, R.; Ramachandran, S. A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. *Genome Biol. Evol.* **2019**, *11*, 2078–2098. [CrossRef]
25. Krokida, A.; Delis, C.; Geisler, K.; Garagounis, C.; Tsikou, D.; Peña-Rodríguez, L.M.; Katsarou, D.; Field, B.; Osbourn, A.E.; Papadopoulou, K.K. A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.* **2013**, *200*, 675–690. [CrossRef]
26. Matsuba, Y.; Nguyen, T.T.; Wiegert, K.; Falara, V.; Gonzales-Vigil, E.; Leong, B.; Schafer, P.; Kudrna, D.; Wing, R.A.; Bolger, A.M.; et al. Evolution of a complex locus for terpene biosynthesis in solanum. *Plant Cell* **2013**, *25*, 2022–2036. [CrossRef] [PubMed]
27. Huffaker, A.; Kaplan, F.; Vaughan, M.M.; Dafoe, N.J.; Ni, X.; Rocca, J.R.; Alborn, H.T.; Teal, P.E.; Schmelz, E.A. Novel acidic sesquiterpenoids constitute a dominant class of pathogen-induced phytoalexins in maize. *Plant Physiol.* **2011**, *156*, 2082–2097. [CrossRef] [PubMed]
28. Osbourn, A.; Papadopoulou, K.K.; Qi, X.; Field, B.; Wegel, E. Finding and analyzing plant metabolic gene clusters. *Methods Enzymol.* **2012**, *517*, 113–138. [CrossRef] [PubMed]
29. Blin, K.; Medema, M.H.; Kazempour, D.; Fischbach, M.A.; Breitling, R.; Takano, E.; Weber, T. antiSMASH 2.0—A versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **2013**, *41*, W204–W212. [CrossRef]
30. Edwards, K.D.; Fernandez-Pozo, N.; Drake-Stowe, K.; Humphry, M.; Evans, A.D.; Bombarely, A.; Allen, F.; Hurst, R.; White, B.; Kernodle, S.P.; et al. A reference genome for Nicotiana tabacum enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genom.* **2017**, *18*, 448. [CrossRef]
31. Fernandez-Pozo, N.; Menda, N.; Edwards, J.D.; Saha, S.; Tecle, I.Y.; Strickler, S.R.; Bombarely, A.; Fisher-York, T.; Pujar, A.; Foerster, H.; et al. The Sol Genomics Network (SGN)—From genotype to phenotype to breeding. *Nucleic Acids Res.* **2015**, *43*, D1036–D1041. [CrossRef]
32. Falara, V.; Akhtar, T.A.; Nguyen, T.T.; Spyropoulou, 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]
33. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]

34. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef]
35. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2015**, *44*, D457–D462. [CrossRef] [PubMed]
36. Wang, W.F.; Chen, P.; Lv, J.; Chen, L.; Sun, Y.H. Transcriptomic analysis of topping-induced axillary shoot outgrowth in *Nicotiana tabacum*. *Gene* **2018**, *646*, 169–180. [CrossRef] [PubMed]
37. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105–1111. [CrossRef]
38. Russo, P.S.T.; Ferreira, G.R.; Cardozo, L.E.; Burger, M.C.; Arias-Carrasco, R.; Maruyama, S.R.; Hirata, T.D.C.; Lima, D.S.; Passos, F.M.; Fukutani, K.F.; et al. CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinform.* **2018**, *19*, 56. [CrossRef]
39. Kautsar, S.A.; Suarez Duran, H.G.; Blin, K.; Osbourn, A.; Medema, M.H. plantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **2017**, *45*, W55–W63. [CrossRef]
40. Buels, R.; Yao, E.; Diesh, C.M.; Hayes, R.D.; Munoz-Torres, M.; Helt, G.; Goodstein, D.M.; Elsik, C.G.; Lewis, S.E.; Stein, L.; et al. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* **2016**, *17*, 66. [CrossRef]
41. 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]
42. Panter, P.E.; Muranaka, T.; Cuitun-Coronado, D.; Graham, C.A.; Yochikawa, A.; Kudoh, H.; Dodd, A.N. Circadian Regulation of the Plant Transcriptome Under Natural Conditions. *Front. Genet.* **2019**, *10*, 1239. [CrossRef]
43. Usadel, B.; Obayashi, T.; Mutwil, M.; Giorgi, F.M.; Bassel, G.W.; Tanimoto, M.; Chow, A.; Steinhauser, D.; Persson, S.; Provart, N.J. Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. *Plant Cell Environ.* **2009**, *32*, 1633–1651. [CrossRef]
44. Zhou, Q.; Xie, Z.; Zhang, Z.; Cao, Y.; Zhang, J.; Chen, S. Cloning, expression and characterization of a nucleoside diphosphate kinase (NDPK) gene from tobacco. *Prog. Nat. Sci.* **2007**, *17*, 906–912. [CrossRef]
45. De Rosa, A.; Watson-Lazowski, A.; Evans, J.R.; Groszmann, M. Genome-wide identification and characterisation of Aquaporins in *Nicotiana tabacum* and their relationships with other Solanaceae species. *BMC Plant Biol.* **2020**, *20*, 266. [CrossRef]
46. Boutanaev, A.M.; Moses, T.; Zi, J.; Nelson, D.R.; Mugford, S.T.; Peters, R.J.; Osbourn, A. Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E81–E88. [CrossRef]
47. Liu, X.; Zhu, X.; Wang, H.; Liu, T.; Cheng, J.; Jiang, H. Discovery and modification of cytochrome P450 for plant natural products biosynthesis. *Synth. Syst. Biotechnol.* **2020**, *5*, 187–199. [CrossRef] [PubMed]
48. Nguyen, T.-D.; Dang, T.-T.T. Cytochrome P450 Enzymes as Key Drivers of Alkaloid Chemical Diversification in Plants. *Front. Plant Sci.* **2021**, *12*, 682181. [CrossRef] [PubMed]
49. Christianson, D.W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **2017**, *117*, 11570–11648. [CrossRef]
50. Martin, D.M.; Aubourg, S.; Schouwey, M.B.; Daviet, L.; Schalk, M.; Toub, O.; Lund, S.T.; Bohlmann, J. Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **2010**, *10*, 226. [CrossRef]
51. Polturak, G.; Liu, Z.; Osbourn, A. New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. *Curr. Opin. Green Sustain. Chem.* **2022**, *33*, 100568. [CrossRef]
52. Mao, L.; Kawaide, H.; Higuchi, T.; Chen, M.; Miyamoto, K.; Hirata, Y.; Kimura, H.; Miyazaki, S.; Teruya, M.; Fujiwara, K.; et al. Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 12472–12480. [CrossRef] [PubMed]
53. Chen, X.; Luck, K.; Rabe, P.; Dinh, C.Q.; Shaulsky, G.; Nelson, D.R.; Gershenzon, J.; Dickschat, J.S.; Kollner, T.G.; Chen, F. A terpene synthase-cytochrome P450 cluster in *Dictyostelium discoideum* produces a novel trisnorsesquiterpene. *eLife* **2019**, *8*, e44352. [CrossRef] [PubMed]
54. Bathe, U.; Tissier, A. Cytochrome P450 enzymes: A driving force of plant diterpene diversity. *Phytochemistry* **2019**, *161*, 149–162. [CrossRef]
55. Field, B.; Osbourn, A.E. Metabolic Diversification-Independent Assembly of Operon-Like Gene Clusters in Different Plants. *Science* **2008**, *320*, 543–547. [CrossRef]
56. Wisecaver, J.H.; Borowsky, A.T.; Tzin, V.; Jander, G.; Kliebenstein, D.J.; Rokas, A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **2017**, *29*, 944–959. [CrossRef] [PubMed]