Planta
https://doi.org/10.1007/s00425-019-03255-7

ORIGINAL ARTICLE

Characterization and evolution of gene clusters for terpenoid phytoalexin biosynthesis in tobacco

Xi Chen[1,2] · Fangjie Liu[1,2] · Lu Liu[1] · Jie Qiu[1] · Dunhuang Fang[3] · Weidi Wang[1] · Xingcheng Zhang[1] · Chuyu Ye[1] · Michael Paul Timko[4] · Qian-Hao Zhu[5] · Longjiang Fan[1,2] · Bingguang Xiao[3]

Received: 23 April 2019 / Accepted: 6 August 2019

Abstract

Main conclusion **The study performed genome-wide identification, characterization and evolution analysis of gene clusters for phytoalexin terpenoid biosynthesis in tobacco, and specifically illustrated ones for capsidiol, an efficient defensive specialized metabolite.**

Abstract Terpenoid phytoalexins play an important role in plant self-defense against pest and pathogen attack. Terpenoid biosynthesis involves terpene synthase and cytochrome P450, which always locate and function as cluster(s). In this study, we performed genome-wide investigation of metabolic gene clusters involved in terpenoid production in tobacco (*Nicotiana tabacum*). Due to the complexity of the tobacco genome, we modified a published prediction pipeline to reduce the influence of the large number of repeats and to improve the annotation of tobacco genes with respect to their metabolic functions. We identified 1181 metabolic gene clusters with 34 of them potentially being involved in terpenoid biosynthesis. Through integration with transcriptome and metabolic pathway annotation analyses, 3 of the 34 terpenoid biosynthesis-related gene clusters were determined to be high-confidence ones, with 2 involved in biosynthesis of capsidiol, a terpenoid recognized as 1 of the effective resistance compounds in the *Nicotiana* species. The capsidiol-related gene cluster was conserved in *N. sylvestris*, *N. tomentosiformis* and *N. attenuate*. Our findings demonstrate that phytoalexins in tobacco can arise from operon-like gene clusters, a genomic pattern characterized as being beneficial for rapid stress response, gene co-regulation, co-function and co-heredity.

Keywords Capsidiol · Cytochrome P450 · Genome-wide identification · *Nicotiana tabacum* · Terpene synthase · Transcriptome

Abbreviations

CYP	Cytochrome P450
EAH	5-*Epi*-aristolochene dihydroxylase
EAS	5-*Epi*-aristolochene synthase
MGC	Metabolic gene cluster
TPS	Terpene synthase

Xi Chen and Fangjie Liu contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00425-019-03255-7) contains supplementary material, which is available to authorized users.

✉ Bingguang Xiao
xiaobgsubmission@126.com; xiaobg@cpu.edu.cn

Extended author information available on the last page of the article

Introduction

Plant damage caused by pests and pathogen attacks significantly limit worldwide agricultural production. Phytoalexins induced by a pathogen are synthesized de novo and rapidly accumulated and transported, and are an important part of the plant defense repertoire (Ahuja et al. 2012). To date, 32 classes of pathways involved in biosynthesis of phytoalexins have been documented in the plant metabolic network (Zhang et al. 2010). Of these pathways, terpenoid-related pathways are predominant. Diterpenoids, such as the phytoalexins phytocassane, oryzalexin, and momilactone, in rice (*Oryza sativa*) exhibit direct anti-fungal activity against the pathogen *Magnaporthe grisea*, which causes rice leaf blast disease (Peters 2006; Shimura et al. 2007). Momilactone A also contributes to anti-herbivore activity against *Sogatella furcifera* in rice (Kanno et al. 2012). Sesquiterpenoid

Published online: 14 August 2019

phytoalexins such as zealexin in maize (*Zea mays*), rishitin in the *Solanum* species, capsidiol in *Nicotiana* and the *Capsicum* species are important inducible defensive compounds (Huffaker et al. 2011). The biosynthesis of terpenoids mostly involves two classes of enzymes, terpene synthase (TPS) and cytochrome P450 (CYP). Studies have revealed that genes encoding TPS and CYP are inclined to locate and function as so-called metabolic gene clusters (MGCs) (Field and Osbourn 2008; Matsuba et al. 2013; Boutanaev et al. 2015). For example, in rice, the momilactone A biosynthesis pathway includes TPS-coding genes *CPS4* and *KSL4*, and CYP-coding genes *CYP99A2* and *CYP99A3*. These four gene clusters are within a genomic region of ~ 180 kb on chromosome 4 (Shimura et al. 2007). The core genes, including *CPS2*, *KSL5*, *KSL6*, *KSL7* (for TPS) and *CYP71Z6*, *CYP71Z7*, *CYP76M7*, *CYP76M8* (for CYP), involved in biosynthesis of phytocassane/oryzalides form a cluster within a ~ 250-kb genomic region on chromosome 2 (Swaminathan et al. 2009). On the other hand, it was also reported that gene clusters were often formed by pairing between genes encoding TPS and CYP (Boutanaev et al. 2015).

The most salient feature of the genuine MGCs is the existence of genes that encode signature enzymes that synthesize the scaffold of the specialized metabolites and tailoring enzymes that modify the scaffold in its various chemical groups to form the end-product (Boycheva et al. 2014). Hallmark signature enzymes include terpene synthases (TPS), phenylpropanoid signature enzymes (PSE), alkaloid signature enzymes (ASE) and polyketide synthases (PKS). Tailoring enzymes include cytochrome P450 (P450), 2-oxoglutarate-dependent dioxygenases (2ODD) and methyltransferase, acyltransferase, and glycosyltransferase. Genes encoding these enzymes arranged in the co-occurrence model greatly contribute to rapid stress response, gene co-regulation, co-function and co-heredity, in a manner similar to the operon in prokaryotic genomes (Boycheva et al. 2014). Identification of the functional MGCs in crops will dramatically improve the efficiency of genome editing for improving economical traits.

To uncover the most reliable MGCs from plant genomes, two state-of-the-art in silico approaches have recently been developed. One is the web-based tool plantiSMASH (http://plantismash.secondarymetabolites.org) (Kautsar et al. 2017), and the other is the algorithm published by the Schlapfer group for gene cluster identification (Schlapfer et al. 2017). The plantiSMASH pipeline employs a comprehensive library of profile Hidden Markov models (pHMMs), including 62 enzyme families known to be involved in plant biosynthetic pathways, and the CD-HIT clustering algorithm for gene cluster prediction. PlantiSMASH has been used to identify 2007 MGCs from more than 47 plant genomes, and 380 of these MGCs were related to terpene synthesis. PlantiSMASH is built on a flexible and user-friendly platform, and compatible with multiple input file formats. However, plantiSMASH only provides the core domains of the MGC genes, and is constrained by the 62 enzyme families, which largely limits its mining power for gene clusters. Schlapfer et al. (2017) presented a more comprehensive computational pipeline for MGC prediction, including identification of metabolic enzymes (a machine learning-based tool, E2P2), pathways (the Pathway Tools), and gene clusters (PlantClusterFinder), to generate a much larger number of enzyme entries and provide more information for the annotated genes. Schlapfer's group has detected 11,969 gene clusters from 18 plant species. Distinct from plantiSMASH which employed HMMER for PFAM domain identification, PlantClusterFinder relies on the organism-specific Pathway/Genome Databases (PGDBs) for MGC identification and annotation, which integrates abundant metabolism resources. However, the plentiful metabolism resources increase prediction sensitivity at the cost of false positives.

Tobacco (*Nicotiana tabacum*) produces thousands of specialized metabolites and some, especially terpenoids, are used as phytoalexins (Jassbi et al. 2017). For example, diterpene alcohols (30 isolated) and glycosides (14 isolated) play an important defensive role against aphids (Jassbi et al. 2017). To date, more than 29 monoterpenoids and 85 sesquiterpenoids have been reported in tobacco as phytoalexins or feeding deterrents against pathogens or herbivores (Jassbi et al. 2017). Capsidiol ($C_{15}H_{24}O_2$, 236.35 Da), one of the sesquiterpenoid phytoalexins, was first isolated from TMV (*Tobacco Mosaic Virus*)-infected leaves of *N. tabacum* and exhibited antifungal activity against *Cladosporium cucumerinum*, *Phytophthora infestans*, and *Potato virus X* (Bailey et al. 1975; Matsukawa et al. 2013; Li et al. 2015). Capsidiol synthesis requires two enzymes: 5-*epi*-aristolochene synthase (EAS) and 5-*epi*-aristolochene dihydroxylase (EAH), and three reaction steps (Fig. 1) (Starks et al. 1997; Ralston et al. 2001). EAS is a kind of TPS that catalyzes farnesyl diphosphate (FPP) to 5-*epi*-aristolochene, while EAH also termed CYP71D20 is one member of the CYP450 clan 71, and responsible for the conversion of 5-*epi*-aristolochene to its dihydroxylated form, capsidiol (Fig. 1). Lee et al. (2017) proved the presence of two gene clusters for capsidiol biosynthesis in pepper and claimed that capsidiol enhanced resistance to non-adapted pathogen *P. infestans*, which causes potato late blight disease (Lee et al. 2017). The association of terpene-derived phytoalexins and MGC in tobacco has not been revealed due to the complexity of the tobacco genome. Fortunately, the tobacco genome has been updated recently (Edwards et al. 2017). Sierro et al. (2014) presented the first tobacco genome (cultivar K326) with a ~ 3.7-Gb assembled sequence (Sierro et al. 2014). Edwards et al. (2017) improved the assembly by increasing the assembly to ~ 4.5 Gb with the scaffolds organized

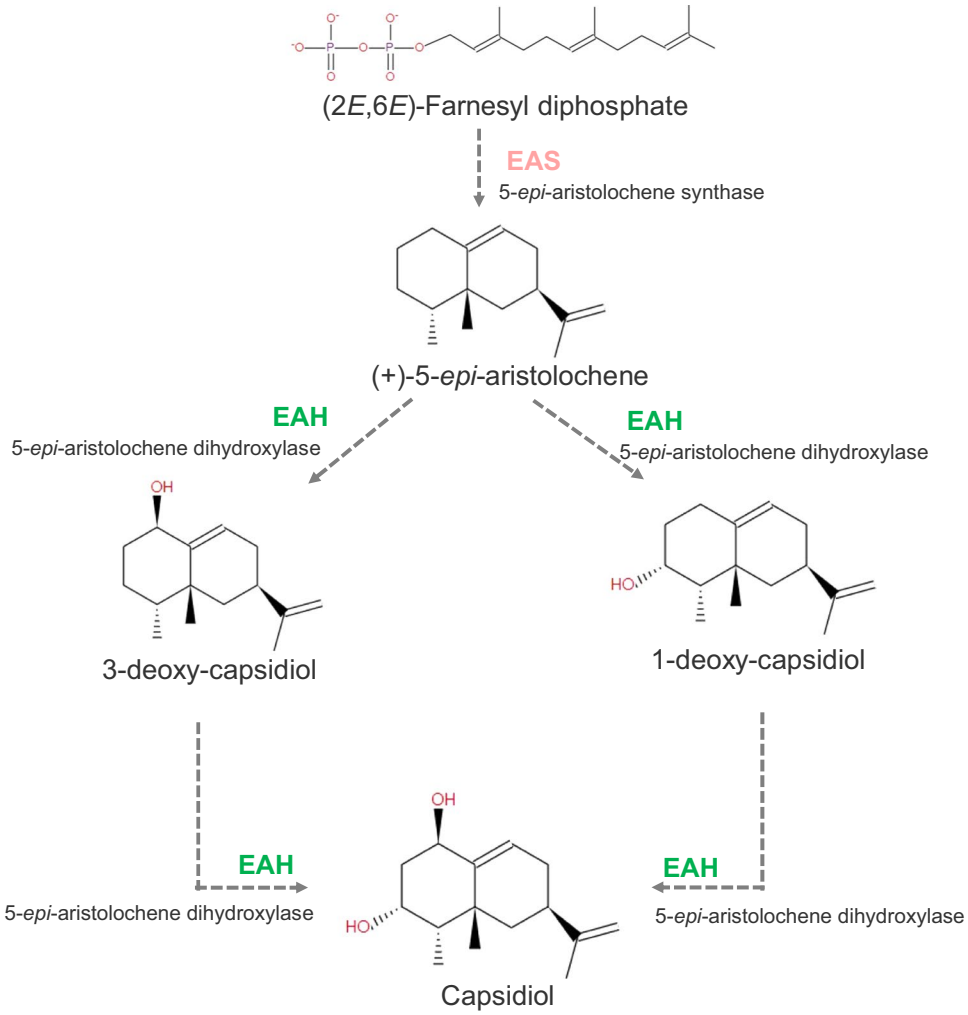


Fig. 1 Diagram of the capsidiol biosynthetic pathway (PWY-2921, MetaCyc). Two enzymes (5-*epi*-aristolochene synthase or EAS and 5-*epi*-aristolochene dihydroxylase or EAH) and three steps of reactions are responsible for capsidiol biosynthesis. EAS catalyzes FPP to 5-*epi*-aristolochene, EAH then hydroxylases 5-*epi*-aristolochene at C-1 or C-3 to form capsidiol by two-step hydroxylation

into 24 pseudomolecules. Even though the allotetraploid *N. tabacum* ($2n = 4x = 48$) stands out due to its big genome and significant proportion (~ 70%) of repeats, the availability of a large amount of the genome resource and the improved genome assembly encouraged us to mine MGCs in tobacco.

To comprehensively and credibly investigate high-confidence MGCs involved in metabolic pathways in tobacco, we first performed MGC prediction with a workflow modified from PlantClusterFinder and then re-detected by plantiSMASH. Overall, 1181 putative MGCs were predicted with 34 of them potentially involved in terpenoid biosynthesis. Three MGCs were found to be involved in biosynthesis of capsidiol. According to transcriptome analysis, the core gene pairs *EAS–EAH* in MGCs involved in capsidiol biosynthesis showed a highly significant co-expression pattern. Collinearity analysis showed the co-occurrence of the capsidiol biosynthesis MGCs in tobacco's ancestors, *N. sylvestris* and *N. tomentosiformis*. The MGCs identified in this study provide a possibility and a novel perspective to functionally investigate the clustered metabolic pathway involved in biosynthesis of terpene-derived phytoalexins in tobacco.

Materials and methods

Genomic and annotation data

The genomes and the annotation files used in this study were downloaded from Sol Genomics Network (SGN, https://solgenomics.net/), including genome sequences of *N. tabacum* cultivar K326 (Edwards et al. 2017), *N. tabacum* cultivar TN90 (Sierro et al. 2014), *N. tabacum* cultivar BX (Sierro et al. 2014), *N. sylvestris* (Sierro et al. 2013), and *N. tomentosiformis* (Sierro et al. 2013), genome release v2 and annotation v5 of *N. attenuata* (Xu et al. 2017), and genome and annotation release v1.55 of *Capsicum annuum* cultivar CM334 (Kim et al. 2014).

Identification of putative enzymes was performed by E2P2 v3.0, downloaded from https://dpb.carnegiescience.edu/labs/rhee-lab/software (Schlapfer et al. 2017). The

annotated protein sequences were used as the input file. The outputs (in .pf format) included the protein ID and the corresponding enzyme category which was classified according to the predicted catalytic functions (Chae et al. 2014).

RNA-seq data

Two sets of RNA-seq data were used in this study. One was downloaded from NCBI and generated from a long day diurnal time-course treatment experiment project including three tissues (root, shoot and shoot apex) (SRP101432; Supplementary Table S1). The other one was from an unpublished pathogen infection experiment done by Yunnan Academy of Tobacco Agricultural Sciences. In this experiment, RNAs from two tissues (root and shoot) were sequenced. Raw reads were first filtered to keep clean data using NGSQC toolkit v2.3.3 with the default settings (Patel and Jain 2012). The clean reads were then aligned to the K326 genome assembly using HISAT v2.1.0 (Pertea et al. 2016). After alignment, the number of reads mapped to each predicted transcript in each sample was counted and normalized to fragments per kilo base of exon per million fragments (FPKM) using StringTie v1.3.4d (Pertea et al. 2016).

Metabolic gene cluster prediction and classification

Prediction of MGCs was performed by a modified method (Fig. 2) based on Schlapfer et al. (2017). Briefly, the E2P2 output file was used as input of Pathway Tools, a state-of-the-art software for specie-specific metabolic pathway database construction. Taxonomic range-based (NCBI-TAXON-ID 4097) pathway inference and pathway database construction were performed by the Pathway Tools' PathoLogic software with the default setting (Karpe et al. 2011), we used the annotation information of the constructed database straightly without validation with the SAVI pipeline to keep as much details as possible for later selection. The generated database library files were then exported for later "co-pathway" filtering.

The shell-based software PlantClusterFinder v1.0 (https://dpb.carnegiescience.edu/labs/rhee-lab/software) (Schlapfer et al. 2017), which was based on sliding window searching, was applied to identify groups of metabolic genes that are contiguously located on the same scaffold. The default parameters were used in running the software except the "Gap Size End", for which we used 10 that was estimated based on the genome feature of tobacco. "Gap Size End" was used to define the maximal number of non-metabolic genes allowed between two genes encoding metabolic enzymes in MGC searching.

MGCs were classified into four subdomains, i.e., terpenoids, phenylpropanoids, alkaloids, and polyketides, based on the metabolites they produce or metabolize. An enzyme library of these metabolites was used in the identification pipeline. In addition, we also included cytochrome P450, 2-oxoglutarate-dependent dioxygenases and methyltransferase, acyltransferase and glycosyltransferase in the library.

High-confidence metabolic gene cluster identification

To acquire a set of high-confidence MGCs among the vast candidates, two kinds of stringent filters were applied. (1) Co-pathway: based on the library file of the metabolic pathway database. We defined co-pathway MGC criteria as: (a) at least two genes within a MGC cluster with a same pathway ID (MetaCyc pathway identifier); (b) two genes should be classified into different reactions (MetaCyc reaction identifier). (2) Co-expression: RNA-seq data were utilized to identify co-expressed genes in each candidate cluster based on two criteria : (a) at least a pair of genes were significant co-expressed (P value ≤ 0.05, under Pearson's product-moment correlation); (b) at least one pair's PCC (Pearson's correlation coefficient, Usadel et al. 2009) over the 90th percentile of the distribution of the genome-wide metabolic gene pairs' PCC (Supplementary Dataset S1, Supplementary Fig. S1). In-house python scripts were used in selection of the MGCs based on the results from the PlantClusterFinder v1.0 pipeline.

Metabolic gene cluster re-detection

PlantiSMASH (http://plantismash.secondarymetabolites.org/) is an online computational tool for MGC prediction (Kautsar et al. 2017). Genome sequence and annotation files in GFF3 format were acquired as inputs. The default parameters were applied in gene cluster re-detection in the K326 genome.

TPS and CYP gene investigation and subfamily assignment

Members of the TPS (PF01397, PF03936) and CYP (PF00067) gene families were identified by InterProScan v5 (Zdobnov and Apweiler 2001). CYP categories have been provided by the SGN annotation file. For TPS subfamily assignment, we downloaded the protein sequences from dicots with known subfamilies (Boutanaev et al. 2015), and the phylogenetic tree was constructed via FastTree (Price et al. 2009) based on protein sequence alignments generated by MAFFT v7.305b (Katoh et al. 2002) using the default settings.

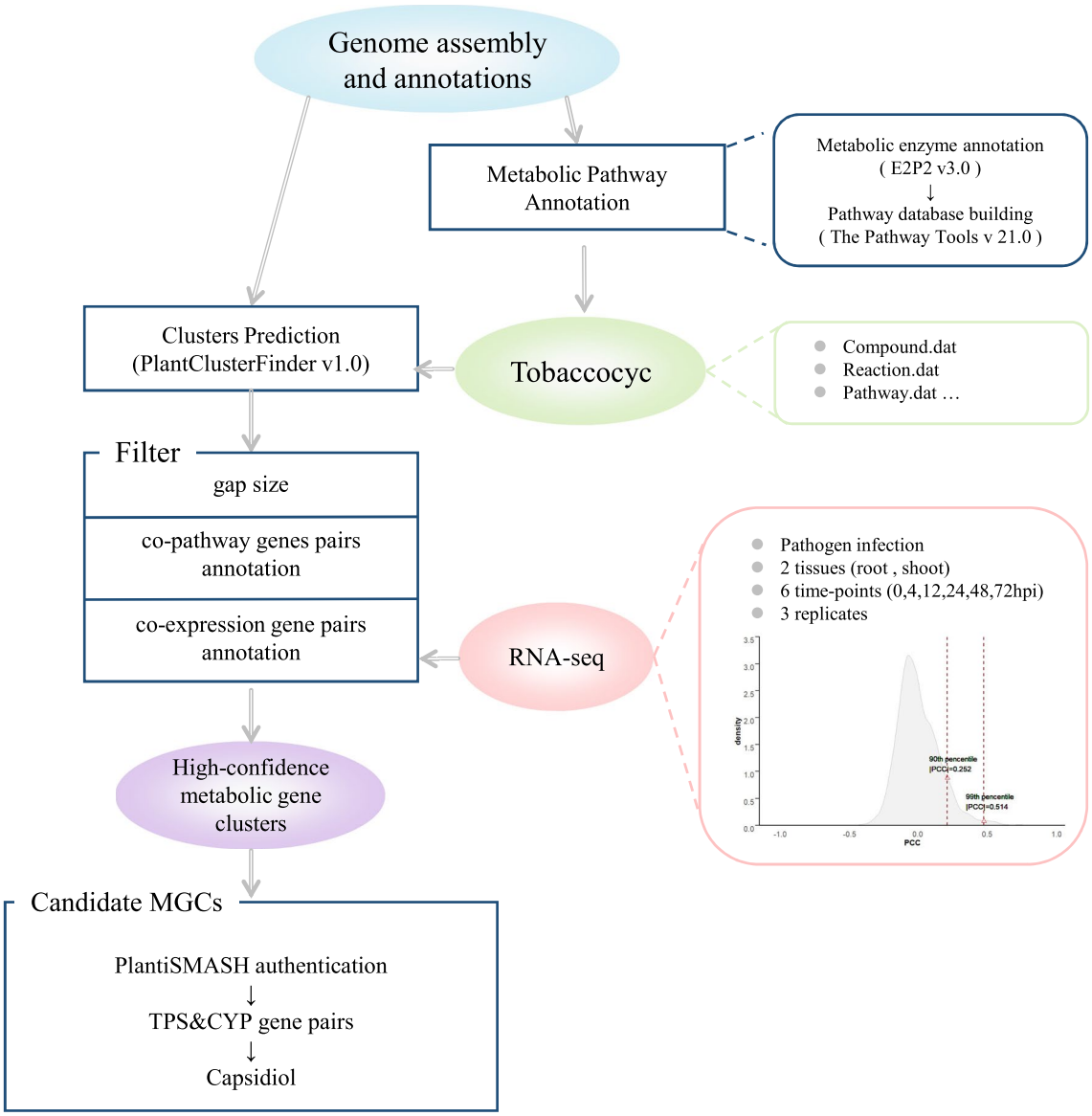


Fig. 2 The pipeline for prediction of metabolic gene clusters. Protein sequences were processed by E2P2 to identify putative enzymes that were then assigned to reactions using the pathway tools (Karpe et al. 2011) and to predict pathways in tobacco, and finally the files for the entire database library were exported under the name of "Tobaccocyc", which was used as the input file for the PlantClusterFinder pipeline. The output of the pipeline was then screened by three filters, i.e., optimal gap size, co-pathway annotation and co-expression annotation, to generate a set of high-confidence MGCs. For the co-expression filter, we used RNA-seq data from a pathogen infection experiment and a previously published experiment (SRP101432), and calculated Pearson's correlation coefficient (PCC) for all the metabolic gene pairs to select MGCs using the 90th percentile PCC as the threshold. After re-detection using plantiSMASH, we obtained MGCs with TPS&CYP gene pairs that were involved in the biosynthesis of capsidiol

TPS and CYP gene pair distribution analysis

As described by Boutanaev et al. (2015), we used the random number generator based on in-house perl scripts to simulate random distribution of gene pairs on the whole-genome scale. The average count of 1000 computer simulation was compared with the observed distributions of *TPS&CYP* gene pairs with a distance less than or equal to 30, 50, 100, 150 or 200 kb from each other (Supplementary Fig. S2), since these intervals were considered close proximity for gene pairs to be associated (Boutanaev et al. 2015). To test whether the TPS&CYP pairing's near locating happened by chance, χ^2 tests were applied across all these five intervals to compare the observed and random count of TPS/CYP pairs by custom R scripts, and followed with the calculation process:

$$\chi^2 = \sum_{\text{interval}}^{\{30,50,100,150,200\}} \frac{(\text{observed} - \text{random})^2}{\text{random}} \quad (df = 4).$$

Identification of orthogroups

With the protein sequence files from *N. tabacum* cultivar K326, *N. sylvestris*, *N. tomentosiformis*, *N. attenuate*, *N. tabacum* cultivar TN90, *N. tabacum* cultivar BX, *C. annuum* (downloaded from Sol Genomics Network), orthogroups were identified among these species using OrthoFinder v2.2.7 (Emms and Kelly 2015) with the default settings (BLASTP *E* value ≤ 1e−5 and MCL inflation parameter of 1.5).

Results

Genome-wide prediction of metabolic gene clusters in common tobacco

Following the prediction pipeline shown in Fig. 2 (for details, see "Materials and methods"), we created a pathway database with 11,789 enzymes, 4408 reactions and 557 metabolic pathways for the tobacco cultivar K326, which was then used in MGC annotation and determination. Compared with the dataset of *N. tabacum* cultivar TN90 on the plant metabolic network (https://plantcyc.org/databases/ntabacumtn90cyc/2.0), which includes 9155 enzymes, 3146 reactions and 506 metabolic pathways, the main difference may come from the features of the genome and the assembly annotation quality. As a result, 1181 MGCs involving 9071 genes were predicted, with 4884 being metabolic genes (Table 1). The predicted MGCs had a physical size ranging from 4 to 1161 kb, with a median size of ~ 163 kb (Fig. 3a). Of the MGCs, 13% contained seven to eight genes which represented the largest portion of the identified putative MGCs (Fig. 3c), with four to five being metabolic genes (Fig. 3b). The predicted size and number of genes of the tobacco MGCs are consistent with those of previously experimentally verified MGCs in other plants (33–284 kb with 4–18 genes) (Schlapfer et al. 2017). Out of the 1181 predicted MGCs, 116 with hallmarks of the mainstream signature genes were classified as specialized MGCs (Fig. 3f, Table 1). Of the 116 specialized MGCs, 32 contained 1–4 tailoring enzymes, mainly CYP450. The predominant products of these predicted specialized MGCs are alkaloid, phenylpropanoid and terpenoid (Fig. 3e, Supplementary Table S2). To verify the prediction results, we used the web-based tool plantiSMASH (Kautsar et al. 2017) with the same tobacco assembly, and in total, 51 MGCs were predicted and 10 were assigned to the terpene class. Forty-one out of the 51 were repeatedly detected, and 5 were assigned to the terpene class (Supplementary Dataset S2).

Co-expression analysis, a powerful strategy for screening genuine gene sets involved in the production of a specialized metabolite, was performed as one of the prerequisites for inferring high-confidence MGCs. We used both the *P* value and PCC threshold in determining significant co-expression patterns (for details, see "Materials and methods"). Integrating the co-expression analysis with metabolic pathway annotation, seven MGCs stood out and were characterized as the high-confidence MGCs for specialized metabolite biosynthesis (Supplementary Fig. S3). Among the seven, two were re-detected by plantiSMASH and both were assigned to the terpene class. The physical sizes of these seven high-confidence MGCs were from 28 kb to 496 kb, with an average of ~ 249 kb (Supplementary Fig. S3). According to the pathway annotation, these seven clusters were involved in sucrose degradation V (sucrose α-glucosidase, PWY66-373), phenylethanol biosynthesis (PWY-5751), methyl indole-3-acetate interconversion (PWY-6303) or capsidiol biosynthesis (PWY-2921). Referring to the signature enzyme types, three out of the seven were classified as the TPS type. All three TPS-type MGCs were predicted to be involved in capsidiol biosynthesis. Signature and tailoring genes in all of the three MGCs were co-up-regulated in the pathogen-infected samples.

Table 1 Overview of the predicted metabolic gene clusters in tobacco

Filters	MGCs	Total gene	Metabolic gene	Specialized	Terpenoid	plantiSMASH
Gap-size (5)	1181	9071	4884	116	34	41
Co-expP.	866	7111	3780	85	27	30
Co-expPCC	454	4117	2176	50	15	20
Co-pwy&expPCC	11	80	57	7	3	2

Numbers of the total MGCs (column 'MGCs'), all and metabolic genes involved (column 'Total gene' and 'Metabolic gene'), specialized MGCs (column 'Specialized'), MGCs involved in terpenoid biosynthesis (column 'Terpenoid'), MGCs re-detected by plantiSMASH (column 'plantiSMASH') are presented. Four screening conditions were investigated, notes are as followed. Notes: "gap-size(5)", MGCs under optimal gap-size 5, which was estimated by software PlantClusterFinder; "co-expP. ", MGCs with at least one gene pair's Pearson correlation test *P* value ≤ 0.05; "co-expPCC", MGCs with at least one gene pair's Pearson correlation coefficients (PCC) ≥ PCC threshold (90th percentile); "co-pwy & expPCC", MGCs with at least one gene pair co-pathway annotated and co-expression pattern over the PCC threshold

Diversification and conservation of metabolic gene clusters in Solanaceae

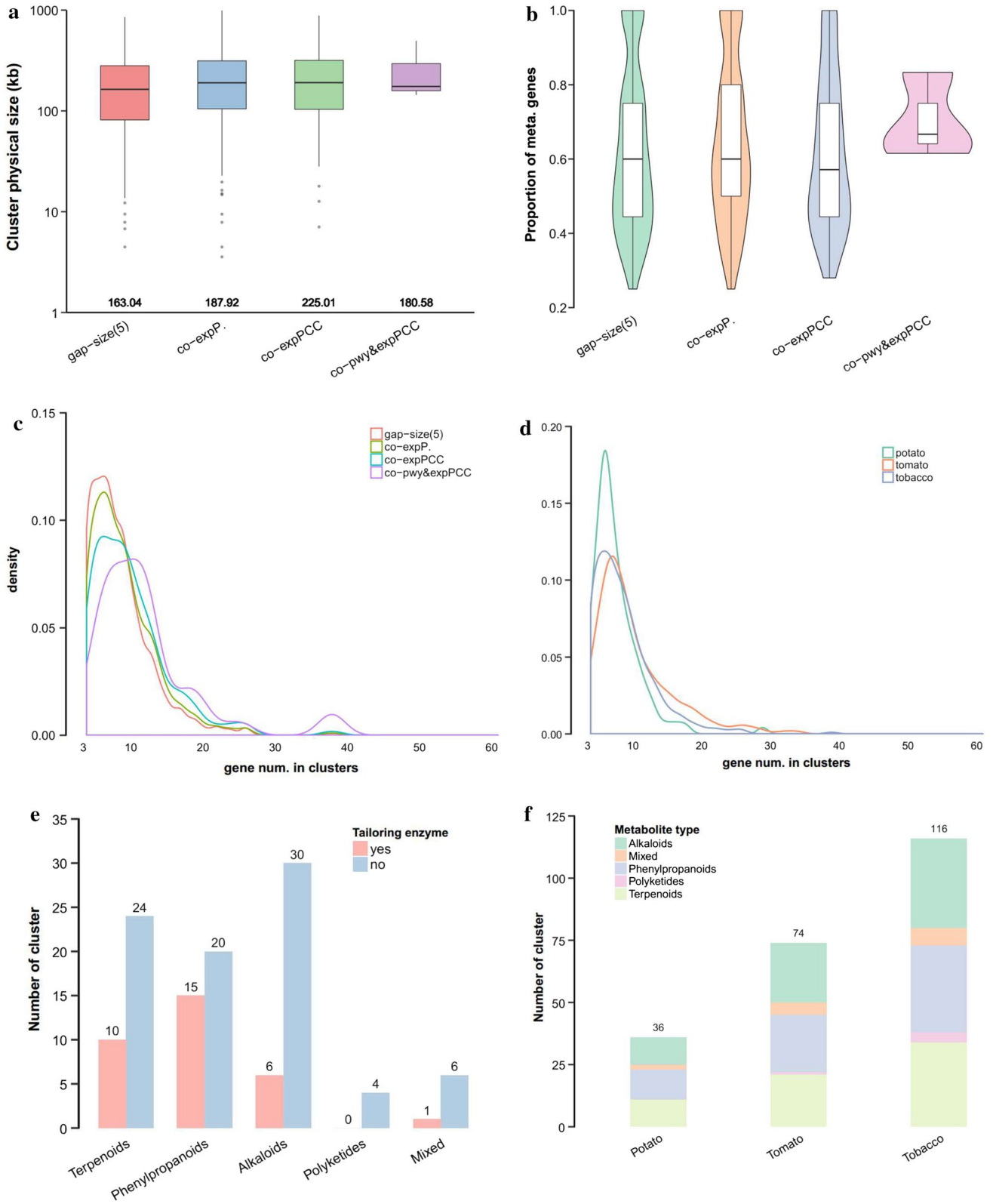
To investigate the MGC diversification and conservation in the Solanaceae family, we compared tobacco MGCs identified here with those previously identified in tomato (*Solanum lycopersicum*; 710 MGCs) and potato (*Solanum tuberosum*; 379 MGCs) (Schlapfer et al. 2017; Supplementary Table S3). Although tobacco has a much larger genome size than both tomato and potato, and is well known for its abundance of secondary metabolites; the number of the predicted MGCs (under optimal gap size) in these three species has no significant difference. However, the physical size of the clusters in tobacco (median size: 163 kb) was much larger than that of the other two species (potato: 35 kb; tomato: 61 kb); Nevertheless, the MGCs from the three species contained a similar number of genes (Fig. 3d). We found 36 and 74 specialized MGCs in potato and tomato, respectively (Fig. 3f, Supplementary Table S4) based on the presence of signature enzymes in the MGCs. These numbers were lower than the 116 identified in tobacco. Theoretically, MGCs with at least two distinct reaction IDs or enzyme IDs in common can be considered to be conserved. Based on this criterion, five conserved MGCs were found in the three species and they were further supported by co-expression analysis. Interestingly, two MGCs (SCF_0003714_150336_234000, SCF_0004447_17210_141828) with conserved reactions were involved in the biosynthesis of morphine (PWY-5270) which has only been reported in *Papaver somniferum* to date (Caspi et al. 2018).

To further characterize the conservation of the MGCs in tobacco, we investigated the presence of a well-characterized potato and tomato gene cluster in tobacco. The cluster is involved in biosynthesis of steroidal glycoalkaloids (potato: α-solanine/α-chaconine, tomato: α-tomatine) (Itkin et al. 2013). α-Solanine and α-tomatine are a kind of steroidal glycoalkaloids and well known for their anti-nutritional substances in solanaceous food plants (tomato, potato and eggplant). In both tomato and potato, two copies of the gene cluster were found (Itkin et al. 2013). We also identified two copies of MGCs annotated to be related to the α-solanine pathway (PWY-5666), one located on the scaffold Nitab 4.5_0001073 spanning a 533.5-kb genomic region with 10 genes, the other spanning a 59.6-kb region with three genes on the scaffold Nitab 4.5_0000170. Based on the OrthoFinder pipeline (Emms and Kelly 2015), we identified six orthologs in the two tobacco MGCs encoding the key enzymes of the α-solanine biosynthesis pathway (Supplementary Fig. S4). α-Solanine biosynthesis requires genes encoding uridine 5′-diphosphate (UDP)-glycosyltransferases to decorate the steroidal alkaloid skeleton with various sugar moieties (Itkin et al. 2013). *GAME1*, *GAME2* in tomato and *SGT1, SGT3* in potato are the genes encoding glycosyltransferase, and their homologs in tobacco were *Nitab 4.5_0001073g0030*, *Nitab 4.5_0000170g0490*, *Nitab 4.5_0001073g0020*, and *Nitab 4.5_0000170g0500*, respectively. Genes encoding CYP450 and dioxygenase are also required for α-solanine biosynthesis, and they were found in the MGC located on scaffold Nitab 4.5_0001073. These genes were co-expressed (*P* value ≤ 0.05) with the four aforementioned genes encoding glycosyltransferase (Supplementary Dataset S3). Thus, equipped with all genes necessary for α-solanine synthesis in the same MGC and co-expressed with genes encoding glycosyltransferases, the MGC on the scaffold Nitab 4.5_0001073 seemed to be a genuine and complete gene cluster for α-solanine biosynthesis in tobacco.

Non-random distribution of TPS and CYP gene pairs for terpenoid biosynthesis in *Nicotiana* genomes

TPS&CYP pairing's non-random distribution offers the potential for an associated TPS&CYP gene pairs' discovery, which often refers to the biosynthesis of terpenoids in a gene-clustering pattern (Field and Osbourn 2008; Matsuba et al. 2013; Boutanaev et al. 2015). Members of the terpene synthase and cytochrome P450 gene families were identified in the genomes of *N. tabacum* and three wild tobacco species (*N. sylvestris*, *N. tomentosiformis*, *N. attenuata*) by InterProScan v5 (Zdobnov and Apweiler 2001). Our results showed that in *N. tabacum, N. sylvestris* and *N. attenuata,* the TPS&CYP combination located in close proximity was significantly different from a random occurrence ($\chi 2$ test, *P* value ≤ 0.01; Supplementary Table S5), which indicated that the association between *TPS* and *CYP* genes in the genomes was likely to function collaboratively.

To better understand the pattern of *TPS* and *CYP* pairing in the whole genome of tobacco and the identified MGCs, we classified the 160 identified *TPS* genes into specific subfamilies (Chen et al. 2011) (Supplementary Fig. S5, Supplementary Dataset S4), and the 798 identified CYP genes into clans (categories provided by the genome annotation files). Chen et al. (2011) assigned TPS proteins from seven plant species into seven subfamilies based on their phylogenic relationships and functions of well-characterized TPS, with the subfamilies of TPS-a and TPS-b regarded as angiosperm-specific (Chen et al. 2011). We found that TPS-a was the biggest subfamily in common tobacco, followed by TPS-b (Supplementary Fig. S5). These two subfamilies had 62 *TPS* genes distributed in 34 MGCs (Supplementary Fig. S5, *TPS* genes are labeled the color pink). TPS in 9 of the 34 MGCs were partnered with CYP as a tailoring enzyme (Supplementary Fig. S5). In these nine MGCs, the distance between TPS and CYP ranged from 2.76 kb to 293.35 kb. TPS-a subfamily genes were more likely to pair with the


a
Cluster physical size (kb)
1000
100
10
1
163.04
187.92
225.01
180.58
gap−size(5)
co−expP.
co−expPCC
co−pwy&expPCC
b
Proportion of meta. genes
1.0
0.8
0.6
0.4
0.2
gap−size(5)
co−expP.
co−expPCC
co−pwy&expPCC
c
density
0.15
0.10
0.05
0.00
gap−size(5)
co−expP.
co−expPCC
co−pwy&expPCC
3
10
20
30
40
50
60
gene num. in clusters
d
0.20
0.15
0.10
0.05
0.00
potato
tomato
tobacco
3
10
20
30
40
50
60
gene num. in clusters
e
Number of cluster
35
30
25
20
15
10
5
0
Tailoring enzyme
yes
no
24
10
20
15
30
6
4
0
6
1
Terpenoids
Phenylpropanoids
Alkaloids
Polyketides
Mixed
f
Number of cluster
125
100
75
50
25
0
Metabolite type
Alkaloids
Mixed
Phenylpropanoids
Polyketides
Terpenoids
36
74
116
Potato
Tomato
Tobacco

◂**Fig. 3** The properties of the predicted metabolic gene clusters in tobacco. **a** Distribution of the physical size of the identified MGCs after each filtering step. Numbers above the *X*-axis indicate the median physical size of MGCs. **b** Distribution of the proportions of metabolic gene number within the identified MGCs after each filtering step. **c** Distribution of the density of the total number of genes within the identified MGCs after each filtering step. **d** Distribution of the density of the total number of genes within the clusters identified based on the filter of optimal gap size in three species (potato, tomato and tobacco). **e** Category of the identified specialized MGCs with the hallmark signature genes in tobacco. For each metabolite type, MGCs were separated into two groups, with or without gene(s) encoding tailoring enzyme(s). **f** Distribution of the MGCs with different type of metabolites in three species (potato, tomato and tobacco), based on optimal gap size

CYP71 clan genes (5/9), followed by TPS-e/f and CYP71 pairing (3/9) (Supplementary Fig. S5). These results are consistent with a previous finding that *TPS* genes were predominantly in combination with the CYP71 clan genes in both eudicots and monocots (Boutanaev et al. 2015).

We also analyzed the conservation of the paired TPS and CYP found in the nine MGCs in the three wild *Nicotiana* species based on orthologous information. TPS&CYP pairs found in five MGCs of common tobacco were found to be conserved in at least one of the three wild tobacco species. The distance between the nearest TPS and CYP coding genes were all ≤ 300 kb, except for those found on chromosome 1 of *N. attenuate* (Supplementary Table S6).

The metabolic gene clusters for capsidiol biosynthesis in tobacco

Three (SCF_0010662_16342_44395, SCF_0002717_73786_293382 and SCF_0001461_22519_518611) of the seven high-confidence MGCs with hallmarks of the mainstream signature genes were identified as being involved in capsidiol biosynthesis (Fig. 4a). Expression of *EAH* (*Nitab 4.5_0010662g0040*) of the MGC SCF_0010662_16342_44395 could not be detected in any analyzed RNA-seq samples, and this MGC was then excluded for further analysis. The remaining two MGCs were both confirmed by plantiSMASH. MGC SCF_0001461_22519_518611 had 13 genes, including two EAH (EAH1: *Nitab 4.5_0001461g0070*, EAH2: *Nitab 4.5_0001461g0130*), four EAS (EAS1: *Nitab 4.5_0001461g0050*, EAS2: *Nitab 4.5_0001461g0140*, EAS3: *Nitab 4.5_0001461g0120*, EAS4: *Nitab 4.5_0001461g0100*), while MGC SCF_0002717_73786_293382 had eight genes, including one EAH (EAH3: *Nitab 4.5_0002717g0030*) and five EAS (EAS5: *Nitab 4.5_0002717g0060*, EAS6: *Nitab 4.5_0002717g0070*, EAS7: *Nitab 4.5_0002717g0050*, EAS8: *Nitab 4.5_0002717g0090*, EAS9: *Nitab 4.5_0002717g0100*). The genomic regions flanking each MGC were also searched for the presence of gene(s) encoding metabolic enzyme(s), and the result was negative. EAH1, EAH2 and EAH3 contain 504, 309 and 503 amino acids, respectively, matching with the previous identified EAH (504 aa, UniProt: Q94FM7) in both length and sequence similarity (Supplementary Fig. S6; EAH1: identity = 94.0%; EAH2: identity = 82.6%; EAH3: identity = 86.7%), even though EAH2 was shorter. Similarly, comparing with the known EAS (548 aa, UniProt: Q40577), two of the seven EAS proteins (EAS1 and EAS7) had the same number of amino acids and showed very high similarity (Supplementary Fig. S6; EAS1: identity = 94.3%; EAS7: identity = 96.2%), while the others were shorter. Complete information about these core members are presented in Table 2. SCF_0002717_73786_293382 and SCF_0001461_22519_518611 each contained a pair of full length *EAS* and *EAH* genes (*EAS7* and *EAH3* for SCF_0002717_73786_293382 and *EAS1* and *EAH1* for SCF_0001461_22519_518611), so both MGCs are thus supposed to be functional for capsidiol biosynthesis. Co-expression of the paired *EAS* and *EAH* genes would suggest a functional MGC. We, therefore, investigated the expression levels of the genes in each MGC and their co-expression patterns. Not surprisingly, *EAS1* and *EAH1* were co-expressed in SCF_0001461_22519_518611 (Pearson correlation test PCC = 0.99 and *P* value = 0). For MGC SCF_0002717_73786_293382, two pairs of *EAS* and *EAH* genes (*EAS7* & *EAH3* and *EAS6* & *EAH3*, PCC = 0.67, *P* value = 3.6e−4 and PCC = 0.76, *P* value = 2.0e−5, respectively) were also co-expressed (Fig. 4b, c).

Evolutionary conservation of the metabolic gene clusters for capsidiol biosynthesis

We then did synteny analysis for the MGCs involved in capsidiol biosynthesis to investigate whether they are conserved in the wild *Nicotiana* species. *N. sylvestris* (female) and *N. tomentosiformis* (male) are the ancestors of common tobacco (Sierro et al. 2013, 2014). For the genes of the MGC SCF_0001461_22519_518611, the *N. tomentosiformis* ortholog of *EAS1* was found in a 33.6-kb synteny region on scaffold ASAG01017751.1, and *N. sylvestris* orthologs of *EAS1* and *EAH1* were found in a 130.0 kb synteny region on scaffold KD972877 (Supplementary Fig. S7). To clarify the origin of the capsidiol biosynthesis MGCs in common tobacco, we aligned the sequence of *N. tabacum* scaffolds Nitab 4.5_0001461 and Nitab 4.5_0002717 with the *N. tomentosiformis* scaffold ASAG01017751.1 and *N. sylvestris* scaffold KD97287. The results showed that Nitab 4.5_0002717 was more likely derived from the T genome, although the origin of Nitab 4.5_0001461 could not be inferred. However, Edwards et al. (2017) who released the *N. tabacum* assembly claimed that the chromosome Nt13 where scaffold Nitab 4.5_0001461 is located was

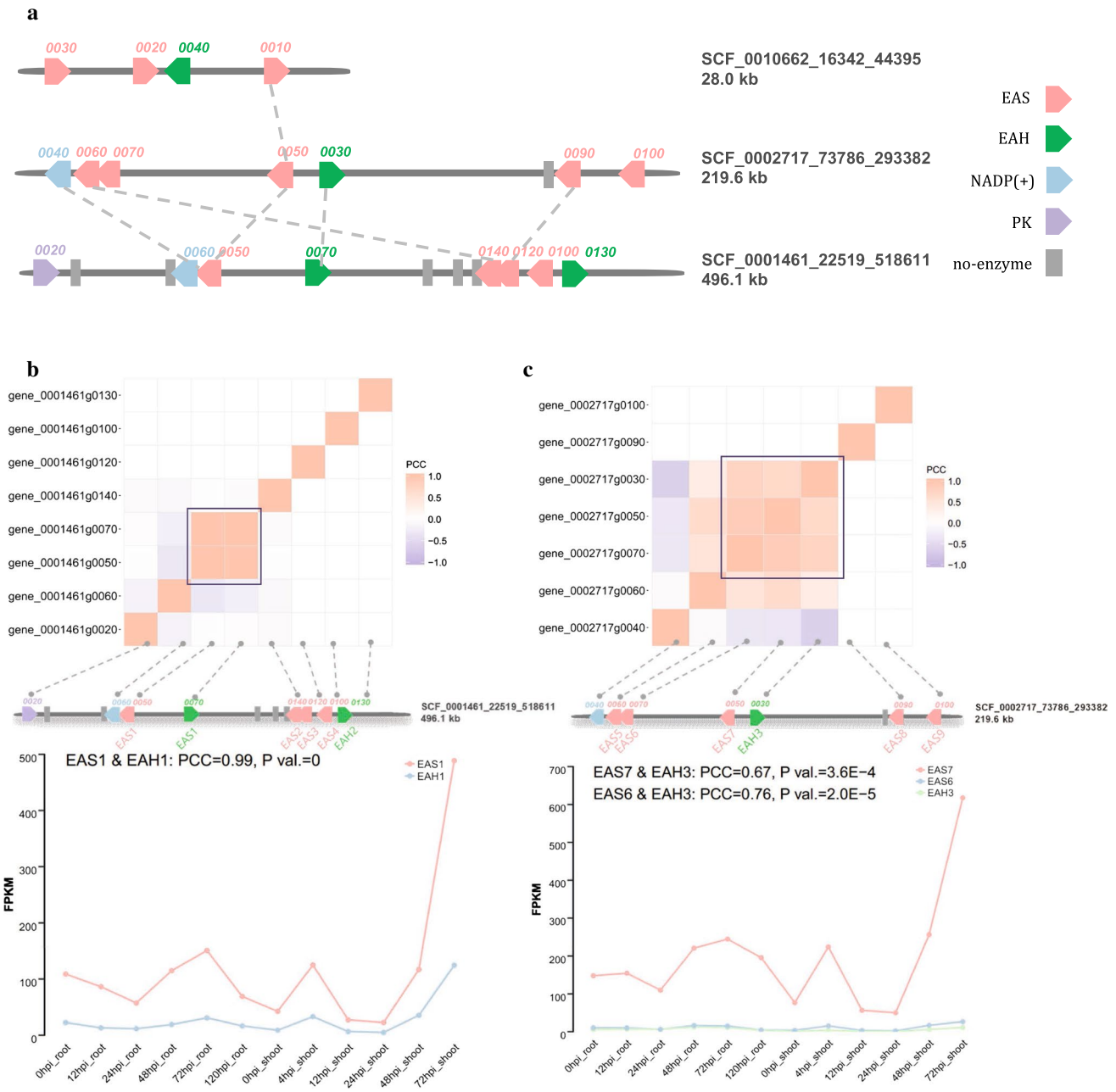


Fig. 4 Functional metabolic gene clusters involved in capsidiol biosynthesis. **a** Genomic structures of the three MGCs involved in capsidiol biosynthesis. Dashed lines indicate the orthologous genes identified by the OrthoFinder pipeline. Due to expression of *EAH* (*Nitab 4.5_0010662g0040*) could not be detected in any analyzed RNA-seq samples, and MGC SCF_0010662_16342_44395 was excluded for further analysis. Genes encoding EAS and EAH showed excellent co-expression patterns in each of the two MGCs. **b** SCF_0001461_22519_518611. **c** SCF_0002717_73786_293382

from the T genome. Therefore, *N. tomentosiformis* seemed to be the donor of both capsidiol biosynthesis MGCs in *N. tabacum*. In the genome of *N. attenuata*, another wild tobacco species, two synteny blocks (8.8 kb and 49.1 Mb) of SCF_0001461_22519_518611 were found, each contained a pair of *EAS* and *EAH* orthologs (Supplementary Fig. S7). Nevertheless, the big synteny block seemed less convincing regarding its length, and the small one was more similar to the MGC on scaffold Nitab 4.5_0002717 based on sequence alignment. We also analyzed SCF_0001461_22519_518611 from cultivar K326 in two other sequences of *N. tabacum* cultivars (TN90 and BX) (Sierro et al. 2014). Each cultivar contained a block (135.5 kb in BX and 238.7 kb in TN90) similar to part of the MGC found in K326 (Supplementary

Table 2 Core members in metabolic gene clusters for capsidiol biosynthesis

Cluster ID	Gene ID	In-house protein ID	Length of protein (aa)	Identity (%)
SCF_0001461_22519_518611	Nitab 4.5_0001461g0070	EAH1	504	94.048
	Nitab 4.5_0001461g0130	EAH2	309	82.593
SCF_0002717_73786_293382	Nitab 4.5_0002717g0030	EAH3	503	86.667
SCF_0001461_22519_518611	Nitab 4.5_0001461g0050	EAS1	548	94.343
	Nitab 4.5_0001461g0140	EAS2	129	84.328
	Nitab 4.5_0001461g0120	EAS3	59	76.316
	Nitab 4.5_0001461g0100	EAS4	292	87.854
SCF_0002717_73786_293382	Nitab 4.5_0002717g0060	EAS5	90	93.976
	Nitab 4.5_0002717g0070	EAS6	230	69.534
	Nitab 4.5_0002717g0050	EAS7	548	96.168
	Nitab 4.5_0002717g0090	EAS8	206	68.571
	Nitab 4.5_0002717g0100	EAS9	310	63.763

Identity (%): results from BLASTP with known EAH (UniProt: Q94FM7) or EAS (UniProt: Q40577)

Fig. S7). Both synteny blocks from BX and TN90 also had a higher sequence similarity with scaffold Nitab 4.5_0002717 from K326.

Capsidiol biogenesis has been well studied in pepper which also belongs to Solanaceae family (Lee et al. 2017). It has been demonstrated that capsidiol is generated by gene clusters in pepper. Besides pepper, common tobacco is the other plant which could produce capsidiol. Therefore, we further extended the synteny analysis to *C. annuum.* By comparing the two MGCs which are responsible for capsidiol biosynthesis in tobacco to that of pepper, we found that the genes and their organization in SCF_0001461_22519_518611 were mostly conserved in *C. annuum* although the length (543.9 kb) of the scaffold Nitab 4.5_0001461 harboring the tobacco MGC was much shorter than the *C. annuum* gene cluster (2.35 Mb) (Fig. 5a). To know whether the synteny region could be further extended in tobacco, we searched for scaffolds containing *EAS* and/or *EAH* genes and adjacency to Nitab 4.5_0001461 based on their chromosome and coordinate information (Edwards et al. 2017). In doing so, we found a scaffold, Nitab 4.5_0004794, which is only 130.8 kb away from Nitab 4.5_0001461 on chromosome 13 (Nt13) and contains two *EAS* genes orthologous to *CA12g05180*, one of the genes from the gene cluster responsible for capsidiol biosynthesis in *C. annuum*; however, no companion *EAH* gene was found in this scaffold (Fig. 5a).

We analyzed the responses of *EAS* and *EAH* genes upon pathogen infection. Of the 27 *EAS* genes, 10 were up-regulated in a variable level, and 5 of them were located on the 3 gene clusters related to capsidiol biosynthesis. In each cluster, a companion *EAH* gene was also up-regulated. The other five up-regulated *EAS* genes were located on five different scaffolds (Nitab 4.5_0000414, Nitab 4.5_0004794, Nitab 4.5_0008226, Nitab 4.5_0013007 and Nitab 4.5_0016435) and without a companion *EAH* gene. Seventeen (6 located on the capsidiol biosynthesis MGCs) of the 27 *EAS* genes showed almost no expression in all samples (Fig. 5b). We also found three up-regulated *EAH* genes without a companion *EAS* gene on three different scaffolds (Nitab 4.5_0020114, Nitab 4.5_0006692, Nitab 4.5_0007566).

Discussion

Here, we performed a genome-wide study of metabolic gene clusters in common tobacco, and identified 1181 candidates, with 34 potentially involved in terpenoid biosynthesis (Table 1, Supplementary Dataset S3). When integrated with transcriptomic and metabolic pathway annotation analyses, we found that two MGCs were involved in capsidiol biosynthesis, and their patterns were conserved in *N. sylvestris*, *N. tomentosiformis* and *N. attenuate* (Fig. 4a, Supplementary Fig. S7). Sesquiterpenoid phytoalexin capsidiol was confirmed to be an effective resistance compound in *Nicotiana* species (Bailey et al. 1975; Li et al. 2015; Jassbi et al. 2017), but the metabolite assembly mechanism for the end-product synthesis has still been unclear. This study demonstrates that the well-known phytoalexins in tobacco can arise from operon-like gene clusters, which are characterized as a genomic pattern benefit for rapid stress response, gene co-regulation, co-function and co-heredity (Boycheva et al. 2014).

In recent years, the clustering of non-homologous and co-localized metabolic genes for biosynthesis of natural products or specialized metabolites in the plant genome has increasingly emerged as an interesting research topic. So far, more than 30 examples of specialized metabolic pathways in plants have been reported to be related to gene clusters (Boycheva et al. 2014; Nutzmann et al. 2016; Guo et al. 2018).

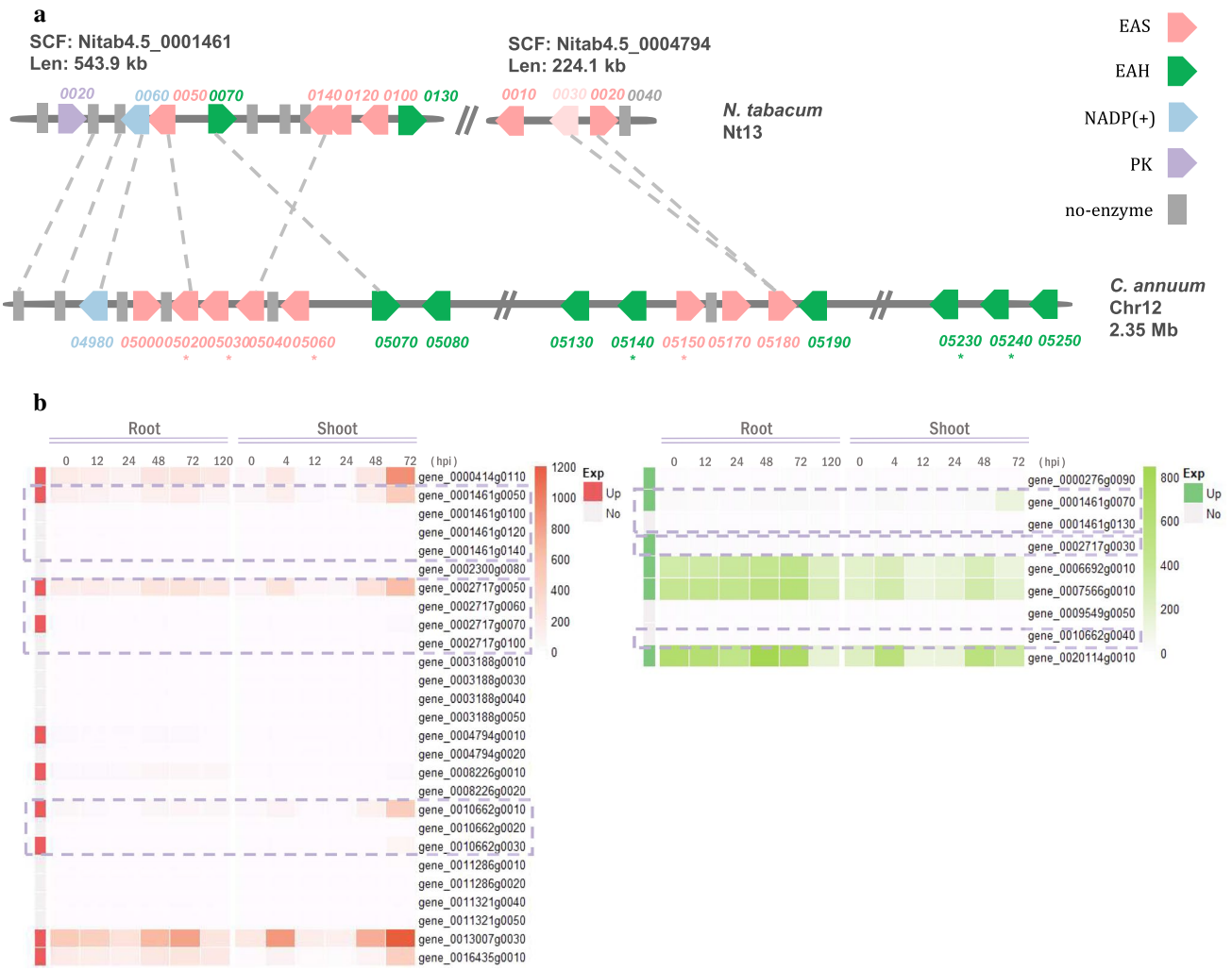


Fig. 5 Potential expansion of the *EAS*/*EAH* beyond the capsidiol metabolic gene clusters. **a** Expansion of the genes encoding EAS and EAH for capsidiol biosynthesis in tobacco and pepper. Collinearity relationships were established based on identification of orthologous genes by the OrthoFinder pipeline. For *C. annuum*, genes labeled with "*" were verified to be significantly highly induced in leaves after pathogen infection. **b** Expression patterns of the 27 genes encoding EAS (EC: 4.2.3.61) and 9 genes encoding EAH (EC: 1.14.13.119) in tobacco root and shoot based on the pathogen infection RNA-seq experiment. RNA-seq was done using 0, 4, 12, 24, 48 and 72 h post-infection (hpi) samples. FPKM was used to represent the expression level. The labels of each gene in the first column were determined by the expression level; "Up" represents that the gene is up-regulated after infection, and "No" means the expression of the gene could not be detected in all analyzed samples, i.e., all FPKM of this gene are equal to 0. Genes appearing in the three detected capsidiol biosynthesis MGCs (SCF_0010662_16342_44395, SCF_0001461_22519_518611 and SCF_0002717_73786_293382) were boxed by dashed lines

The MGCs have a physical size ranging from ~ 33 kb to more than 300 kb, and consist of 3–18 genes (Nutzmann et al. 2016; Schlapfer et al. 2017). Even though it is abundant with bioactive natural products, tobacco has been absent in genome-wide investigation of MGCs.

Given the benefit of the availability of several state-of-the-art in silico approaches for MGC detection, we predicted tobacco MGCs with a modified workflow based on the PlantClusterFinder pipeline and uncovered a vast number of candidates (Fig. 2, Supplementary Dataset S3). While 82% of the MGC results based on plantiSMASH could be recovered by PlantClusterFinder, only 3.5% of MGCs detected by PlantClusterFinder could be re-detected by plantiSMASH, suggesting the candidates predicted by PlantClusterFinder require a further stringent filtering process. We set up filters based on our current knowledge of biosynthesis of metabolites and the expression pattern of MGCs to get 11 high-confidence MGCs from the original 1181 ones. Of these, seven MGCs with hallmarks of signature genes were characterized for specialized metabolite biosynthesis (Supplementary Fig. S3). Among the filters, the two major ones were the co-expression threshold and co-pathway annotation. It is believed that the sufficient

expression data from diverse treatments, developmental stages and tissues were the persuasive evidence for co-expression estimation, however condition-specific correlations caused by diverse samples were difficult to mask (Boycheva et al. 2014; Nutzmann et al. 2016; Schlapfer et al. 2017). Considering that the expression data in our study was not as rich as that of ATTED-II (a plant co-expression database) and due to the complexity and poor annotation of the tobacco genome, we applied the threshold of the 90th percentile PCC distribution of the genome-wide metabolic gene pairs, instead of the 99th one adopted by the PlantClusterFinder pipeline (Supplementary Fig. S1) where the co-expression threshold PCC was inferred by datasets from ATTED-II. Furthermore, because tobacco is very sensitive to treatment, the variances between replicates are hardly controllable, let alone between samples. After applying the 90th percentile of PCC, we still get a relatively higher value of 0.662 compared with the 99th percentile PCC 0.514 for Arabidopsis (Supplementary Fig. S1); and besides, it represented a very high statistical significance (*P* value ≤ 0.001) and a stringent cutoff for co-expression estimation (Supplementary Dataset S1). The other filter we used was co-pathway annotation, which completely relies on the metabolism information provided by the current database. Since the vast majority of the metabolism potential of the plant kingdom still awaits discovery (Nutzmann et al. 2016), merely relying on the current knowledge of metabolism did limit the characterization of genuine MGCs, thus greater effort is required in future research. We defined a co-pathway MGC as having at least one gene pair with a distinct MetaCyc reaction ID and the same pathway ID. This filter would have discarded some genuine MGCs due to the largely unknown co-pathway information of a large number of metabolites in plants. Our filters thus probably have increased the prediction precision at a cost of losing genuine MGCs to some degree. As a result, the number of high-confidence MGCs predicted in this study could be underestimated.

In comparison with the MGCs predicted in potato and tomato, the number of total and specialized MGCs predicted in tobacco is not correlated with its genome size and abundant metabolites (Supplementary Table S3). The quality of gene annotation and genome assembly could be the two key factors affecting the prediction results. By applying the Core Eukaryotic Genes Mapping Approach (Parra et al. 2007), which estimates the completeness of genome annotation, to the tobacco genome we used, tobacco appears to have a complete CEGMA score ~ 85% and a partial CEGMA score ~ 98%, both well over the threshold for building the high-quality metabolic pathway database. Thus, it seems the genome assembly quality would be the reason to explain the low number of MGCs predicted in tobacco. The tobacco assembly we used here is ~ 4.5 Gb, and has only 2% assembly covered by scaffolds with more than 50 genes, which is far away from the criterion (50%) suggested by Schlapfer et al. (2017) for performing MGC prediction. To be qualified as an MGC, a scaffold should contain at least three genes. In tobacco, there is only 57% assembly covered by scaffolds with three or more genes, suggesting that 43% assembly was not used in the prediction (Supplementary Table S3). In addition, a vast majority of the identified MGCs were annotated as being a partial clustering of metabolic pathways, though whether they were actually partial clustering or subjected to the assembly quality deserves further investigation. To conclude, prediction of MGCs in tobacco would be significantly enhanced by improving its genome assembly.

While metabolites are often restricted to specific taxon lineages, comparative genomic analysis among species that share a relative relationship in evolution could still uncover conserved MGCs. In our study, an MGC annotated as involved in the α-solanine biosynthesis pathway exhibited conserved synteny with the α-solanine/α-tomatine gene clusters reported in potato and tomato. In addition, the co-expression pattern of the core component genes was validated by transcriptome analysis (Supplementary Fig. S4). We also discovered the partial clustering of the morphine biosynthesis pathway in tobacco, potato and tomato, and found it to be the only conserved convincing MGC shared by the three plant species. Though morphine biosynthesis has only been reported in *P. somniferum*, the partial clustering of the pathway found in the Solanaceae family may contribute to the production of morphine precursors or other forms of morphine in these species, a topic deserving further study. Furthermore, five out of the nine MGCs with TPS&CYP pairing, including the two key capsidiol biosynthesis MGCs, exhibited conserved synteny with the three wild *Nicotiana* species (Supplementary Table S6). In both the paternal and maternal genomes of common tobacco, we found reliable synteny blocks of MGCs for capsidiol biosynthesis, and sequence analysis suggested that these two tobacco MGCs were likely to originate from *N. tomentosiformis*. Previous studies on repetitive DNA sequence distribution and sequence reads mapping demonstrated a reduced contribution of *N. tomentosiformis* to the tobacco genome (Sierro et al. 2014; Edwards et al. 2017). The well-preserved clustering genomic structure for capsidiol biosynthesis in tobacco and its wild species highlights the significance of capsidiol in the evolution of the *Nicotiana* species. Recently, the gene clustering pattern in a species-specific manner for capsidiol biosynthesis has been reported in pepper (Lee et al. 2017). Based on the collinearity relationship analysis between tobacco and pepper, we discovered potential expansion of *EAS*/*EAH* beyond the capsidiol biosynthesis MGC in pepper (Fig. 5a). However, constrained by the assembly quality of tobacco, whether the two tobacco scaffolds carrying the pathogen-inducible *EAS* or *EAH* are adjacent to each other in the genomic block corresponding to the gene

cluster in pepper is still uncertain, which would compromise our conclusion about the expansion of *EAS*/*EAH* beyond the capsidiol biosynthesis MGC in pepper.

Phytoalexins acting as active weapons against pathogen infection in plants have drawn considerable attention. Our study applied a modified plant cluster prediction strategy to perform a large-scale identification of MGCs in common tobacco. By integrating multiple data analyses (transcriptome and metabolic pathway data sources), two MGCs involved in capsidiol biosynthesis were achieved. As demonstrated in this work, the well-known phytotoxin capsidiol in tobacco was found to be functioning in a clustering pattern, and evolved in a conserved manner among different *Nicotiana* species. The assembly of metabolic pathways within a clustering pattern like an operon is believed to be beneficial for rapid stress response, gene co-regulation, co-function and co-heredity. Our analyses shed light on MGCs' detection in common tobacco with a complex genome, where customized and modified analyses are required. The pipeline raised here could also be used in other plants with complex genomes, such as barley, oilseed rape and cotton. Although our pipeline still has some limitations, we firmly believe that detection of functional MGCs can provide a novel perspective for the investigation of the clustered metabolic pathway, and may greatly contribute to highly efficient genome editing for improving economic traits, especially in disease-resistance breeding.

Author contribution statement XC, BX and FL conceived and designed the research. BX and DF conducted experiments and contributed the RNA-seq data. FL, LL, QJ, WW, XZ and CY analyzed the data. FL and XC wrote the manuscript. LF, MT and QZ revised the manuscript. All the authors read and approved the manuscript.

Acknowledgements This work was supported by the National Natural Science Foundation of China (31860411) to B. Xiao, the Fundamental Research Funds for the Central Universities of China (2017QNA6013) to X. Chen and 111 Project (B17039) to M. Timko.

Data availability Data of the public RNA-seq for tobacco available at SRA Series accession number SRP101432; data of the pathogen infection RNA-seq not publicly available now for our unfinished work of a tobacco pathogen resistance genes cloning project, the data will be released to public in the near future; data of the all the MGCs information available in supplementary material.

References

Ahuja I, Kissen R, Bones AM (2012) Phytoalexins in defense against pathogens. Trends Plant Sci 17:73–90. https://doi.org/10.1016/j.tplants.2011.11.002

Bailey JA, Burden RS, Vincent GGJP (1975) Capsidiol: an antifungal compound produced in *Nicotiana tabacum* and *Nicotiana clevelandii* following infection with tobacco necrosis virus. Phytochemistry 14:597

Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, Osbourn A (2015) Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci USA 112:E81–E88. https://doi.org/10.1073/pnas.1419547112

Boycheva S, Daviet L, Wolfender JL, Fitzpatrick TB (2014) The rise of operon-like gene clusters in plants. Trends Plant Sci 19:447–459. https://doi.org/10.1016/j.tplants.2014.01.013

Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Midford PE, Ong Q, Ong WT, Paley S, Subhraveti P, Karp PD (2018) The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res 46(D1):D633–D639. https://doi.org/10.1093/nar/gkx935

Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized metabolism in plants. Science 344:510–513. https://doi.org/10.1126/science.1252076

Chen F, Tholl D, Bohlmann J, Pichersky E (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J 66:212–229. https://doi.org/10.1111/j.1365-313X.2011.04520.x

Edwards KD, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans AD, Bombarely A, Allen F, Hurst R, White B, Kernodle SP, Bromley JR, Sanchez-Tamburrino JP, Lewis RS, Mueller LA (2017) A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genom 18:448. https://doi.org/10.1186/s12864-017-3791-6

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2

Field B, Osbourn AE (2008) Metabolic diversification–independent assembly of operon-like gene clusters in different plants. Science 320:543–547. https://doi.org/10.1126/science.1154990

Guo L, Qiu J, Li LF, Lu B, Olsen K, Fan L (2018) Genomic clues for crop-weed interactions and evolution. Trends Plant Sci 23:1102–1115. https://doi.org/10.1016/j.tplants.2018.09.009

Huffaker A, Kaplan F, Vaughan MM, Dafoe NJ, Ni X, Rocca JR, Alborn HT, Teal PE, Schmelz EA (2011) Novel acidic sesquiterpenoids constitute a dominant class of pathogen-induced phytoalexins in maize. Plant Physiol 156:2082–2097. https://doi.org/10.1104/pp.111.179457

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, Tikunov Y, Bovy A, Chikate Y, Singh P, Rogachev I, Beekwilder J, Giri AP, Aharoni A (2013) Biosynthesis of antinutritional alkaloids in Solanaceous crops is mediated by clustered genes. Science 341:175–179. https://doi.org/10.1126/science.1240230

Jassbi AR, Zare S, Asadollahi M, Schuman MC (2017) Ecological roles and biological activities of specialized metabolites from the genus *Nicotiana*. Chem Rev 117:12227–12280. https://doi.org/10.1021/acs.chemrev.7b00001

Kanno H, Hasegawa M, Kodama O (2012) Accumulation of salicylic acid, jasmonic acid and phytoalexins in rice, *Oryza sativa*, infested by the white-backed planthopper, *Sogatella furcifera* (Hemiptera: Delphacidae). Appl Entomol Zool 47:27–34. https://doi.org/10.1007/s13355-011-0085-3

Karpe PD, Latendresse M, Caspi R (2011) The pathway tools pathway prediction algorithm. Stand Genom Sci 5:424–429. https://doi.org/10.4056/sigs.1794338

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast

Fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res 45:W55–W63. https://doi.org/10.1093/nar/gkx305

Kim S, Park M, Yeom S-I, Kim Y-M, Lee JM, Lee H-A, Seo E, Choi J, Cheong K, Kim K-T, Jung K, Lee G-W, Oh S-K, Bae C, Kim S-B, Lee H-Y, Kim S-Y, Kim M-S, Kang B-C, Jo YD, Yang H-B, Jeong H-J, Kang W-H, Kwon J-K, Shin C, Lim JY, Park JH, Huh JH, Kim J-S, Kim B-D, Cohen O, Paran I, Suh MC, Lee SB, Kim Y-K, Shin Y, Noh S-J, Park J, Seo YS, Kwon S-Y, Kim HA, Park JM, Kim H-J, Choi S-B, Bosland PW, Reeves G, Jo S-H, Lee B-W, Cho H-T, Choi H-S, Lee M-S, Yu Y, Do Choi Y, Park B-S, van Deynze A, Ashrafi H, Hill T, Kim WT, Pai H-S, Ahn HK, Yeam I, Giovannoni JJ, Rose JKC, Sørensen I, Lee S-J, Kim RW, Choi I-Y, Choi B-S, Lim J-S, Lee Y-H, Choi D (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet 46:270. https://doi.org/10.1038/ng.2877

Lee HA, Kim S, Kim S, Choi D (2017) Expansion of sesquiterpene biosynthetic gene clusters in pepper confers nonhost resistance to the Irish potato famine pathogen. New Phytol 215:1132–1143. https://doi.org/10.1111/nph.14637

Li R, Tee CS, Jiang YL, Jiang XY, Venkatesh PN, Sarojam R, Ye J (2015) A terpenoid phytoalexin plays a role in basal defense of *Nicotiana benthamiana* against *Potato virus X*. Sci Rep 5:9682. https://doi.org/10.1038/srep09682

Matsuba Y, Nguyen TT, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schafer P, Kudrna D, Wing RA, Bolger AM, Usadel B, Tissier A, Fernie AR, Barry CS, Pichersky E (2013) Evolution of a complex locus for terpene biosynthesis in *Solanum*. Plant Cell 25:2022–2036. https://doi.org/10.1105/tpc.113.111013

Matsukawa M, Shibata Y, Ohtsu M, Mizutani A, Mori H, Wang P, Ojika M, Kawakita K, Takemoto D (2013) *Nicotiana benthamiana* calreticulin 3a is required for the ethylene-mediated production of phytoalexins and disease resistance against oomycete pathogen *Phytophthora infestans*. Mol Plant Microbe Interact 26:880–892. https://doi.org/10.1094/MPMI-12-12-0301-R

Nutzmann HW, Huang A, Osbourn A (2016) Plant metabolic clusters—from genetics to genomics. New Phytol 211:771–789. https://doi.org/10.1111/nph.13981

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619. https://doi.org/10.1371/journal.pone.0030619

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11:1650–1667. https://doi.org/10.1038/nprot.2016.095

Peters RJ (2006) Uncovering the complex metabolic network underlying diterpenoid phytoalexin biosynthesis in rice and other cereal crop plants. Phytochemistry 67:2307–2317. https://doi.org/10.1016/j.phytochem.2006.08.009

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. https://doi.org/10.1093/molbev/msp077

Ralston L, Kwon ST, Schoenbeck M, Ralston J, Schenk DJ, Coates RM, Chappell J (2001) Cloning, heterologous expression, and functional characterization of 5-epi-aristolochene-1,3-dihydroxylase from tobacco (*Nicotiana tabacum*). Archives Biochem Biophys 393(2):222–235. https://doi.org/10.1006/abbi.2001.2483

Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T, Kahn D, Rhee SY (2017) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol 173:2041–2059. https://doi.org/10.1104/pp.16.01942

Shimura K, Okada A, Okada K, Jikumaru Y, Ko KW, Toyomasu T, Sassa T, Hasegawa M, Kodama O, Shibuya N, Koga J, Nojiri H, Yamane H (2007) Identification of a biosynthetic gene cluster in rice for momilactones. J Biol Chem 282:34013–34018. https://doi.org/10.1074/jbc.M703344200

Sierro N, Battey JN, Ouadi S, Bovet L, Goepfert S, Bakaher N, Peitsch MC, Ivanov NV (2013) Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome Biol 14:R60. https://doi.org/10.1186/gb-2013-14-6-r60

Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A (2014) The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5:3833. https://doi.org/10.1038/ncomms4833

Starks CM, Back KW, Chappell J, Noel JP (1997) Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. Science 277(5333):1815–1820

Swaminathan S, Morrone D, Wang Q, Fulton DB, Peters RJ (2009) CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. Plant Cell 21:3315–3325. https://doi.org/10.1105/tpc.108.063677

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ 32:1633–1651. https://doi.org/10.1111/j.1365-3040.2009.02040.x

Xu S, Brockmoller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, Lyons E, Pandey P, Pandey SP, Timmermann B, Gaquerel E, Baldwin IT (2017) Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proc Natl Acad Sci USA 114:6133–6138. https://doi.org/10.1073/pnas.1700073114

Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C, Mueller LA, Muller R, Rhee SY (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol 153:1479–1491. https://doi.org/10.1104/pp.110.157396

Affiliations

Xi Chen[1,2] · Fangjie Liu[1,2] · Lu Liu[1] · Jie Qiu[1] · Dunhuang Fang[3] · Weidi Wang[1] · Xingcheng Zhang[1] · Chuyu Ye[1] · Michael Paul Timko[4] · Qian-Hao Zhu[5] · Longjiang Fan[1,2] · Bingguang Xiao[3]

Xi Chen
xich@zju.edu.cn

Fangjie Liu
ajay_liu@zju.edu.cn

Lu Liu
3140101148@zju.edu.cn

Jie Qiu
lefroyqiu@foxmail.com

Dunhuang Fang
fdhkm@sina.com

Weidi Wang
wwd-swxx@foxmail.com

Xingcheng Zhang
113871873@qq.com

Chuyu Ye
yecy@zju.edu.cn

Michael Paul Timko
mpt9g@virginia.edu

Qian-Hao Zhu
Qianhao.Zhu@csiro.au

Longjiang Fan
fanlj@zju.edu.cn

1 Institute of Crop Science and Institute of Bioinformatics, Zhejiang University, Hangzhou 310058, China

2 Research Center for Air Pollution and Health, Zhejiang University, Hangzhou 310058, China

3 Key Laboratory of Tobacco Biotechnological Breeding, Yunnan Academy of Tobacco Agricultural Sciences, Kunming 650021, China

4 Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

5 CSIRO Agriculture and Food, GPO Box 1700, Canberra, ACT 2601, Australia