the plant journal

SEB Society for Experimental Biology

The Plant Journal (2018) doi: 10.1111/tpj.14162

The genome of the medicinal plant *Andrographis paniculata* provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide

Wei Sun[1,†], Liang Leng[1,†], Qinggang Yin[1,†], MeiMei Xu[2,†], Mingkun Huang[1], Zhichao Xu[3], Yujun Zhang[1], Hui Yao[3], Caixia Wang[1], Chao Xiong[1], Sha Chen[1], Chunhong Jiang[4], Ning Xie[4], Xilong Zheng[5], Ying Wang[6], Chi Song[1,‡], Reuben J. Peters[2,‡] and Shilin Chen[1,*,‡]

[1]*Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, China Academy of Chinese Medical Sciences, Institute of Chinese Materia Medica, 100070 Beijing, China,*

[2]*Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011-1079,USA,*

[3]*Key Laboratory of Bioactive Substances and Resources, Utilization of Chinese Herbal Medicine, Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, 100193 Beijing, China,*

[4]*State Key Laboratory of Innovative Natural Medicine and TCM Injections, Jiangxi Qingfeng Pharmaceutical Co. Ltd., 341008 Ganzhou, China,*

[5]*Hainan Branch, Institute of Medicinal Plant Development, 570311 Wanning, China, and*

[6]*Wuhan Benagen Tech Solutions Company Limited, 430070 Wuhan, China*

Received 8 June 2018; revised 29 October 2018; accepted 2 November 2018.
*For correspondence (e-mail slchen@icmm.ac.cn).
†These authors contributed equally to this work.
‡These authors are corresponding co-authors.

SUMMARY

Andrographis paniculata **is a herbaceous dicot plant widely used for its anti-inflammatory and anti-viral properties across its distribution in China, India and other Southeast Asian countries. *A. paniculata* was used as a crucial therapeutic treatment during the influenza epidemic of 1919 in India, and is still used for the treatment of infectious disease in China. *A. paniculata* produces large quantities of the anti-inflammatory diterpenoid lactones andrographolide and neoandrographolide, and their analogs, which are touted to be the next generation of natural anti-inflammatory medicines for lung diseases, hepatitis, neurodegenerative disorders, autoimmune disorders and inflammatory skin diseases. Here, we report a chromosome-scale *A. paniculata* genome sequence of 269 Mb that was assembled by Illumina short reads, PacBio long reads and high-confidence (Hi-C) data. Gene annotation predicted 25 428 protein-coding genes. In order to decipher the genetic underpinning of diterpenoid biosynthesis, transcriptome data from seedlings elicited with methyl jasmonate were also obtained, which enabled the identification of genes encoding diterpenoid synthases, cytochrome P450 monooxygenases, 2-oxoglutarate-dependent dioxygenases and UDP-dependent glycosyltransferases potentially involved in diterpenoid lactone biosynthesis. We further carried out functional characterization of pairs of class-I and -II diterpene synthases, revealing the ability to produce diversified labdane-related diterpene scaffolds. In addition, a glycosyltransferase able to catalyze O-linked glucosylation of andrograpanin, yielding the major active product neoandrographolide, was also identified. Thus, our results demonstrate the utility of the combined genomic and transcriptomic data set generated here for the investigation of the production of the bioactive diterpenoid lactone constituents of the important medicinal herb *A. paniculata*.**

Keywords: genome sequence, *Andrographis paniculata*, diterpenoid, TPS, glucosyltransferase.

INTRODUCTION

Since the dawn of humankind, humans have used botanical medicines for the treatment of ailments (Roy Upton, 2015). There is high demand for herbal medicines in developing countries because of perceived effectiveness, cultural acceptability and affordability (Kamboj, 2000; Pal and Shukla, 2003). Most herbal medicines originated from

© 2018 The Authors
The Plant Journal © 2018 John Wiley & Sons Ltd

traditional therapeutic theory, and the practice can be dated back to before the development and spread of modern medicine. With the advancement of the field of phytochemistry, focus has shifted away from the plant material itself and towards the isolated active compounds (Roy Upton, 2015). Some single-ingredient modern drugs and dietary supplements are derived directly or indirectly from plants, including: artemisinin (from *Artemisia annua* L.), podophyllotoxin (from *Podophyllum peltatum* L.), paclitaxel (from *Taxus brevifolia* Nutt.), vincristine (from *Catharanthus roseus*) and morphine (from *Papaver somniferum* L.) (Raskin *et al.*, 2002; Goel *et al.*, 2008). Meanwhile, prescriptions of herbal medicines involve the simultaneous action of multiple constituents of the plant part as a whole, thought to act synergistically by hitting many molecular targets or just a single target (Li and Weng, 2017). Currently in China both herbal plant material and modern chemically characterized drugs are prescribed.

Andrographis paniculata (Burm. f.) Nees (Figure 1) is a member of the order Lamiales, Acanthaceae (family), widely distributed and used in tropical and subtropical regions of Asia, including India, China, Thailand and Malaysia (Lim *et al.*, 2012). Its use across this region is well known and historically documented in both traditional Chinese medicine (TCM) and the Ayurvedic system of medicine in India (Anju *et al.*, 2012; Pholphana *et al.*, 2013). It is known as 'Chuanxinlian' in Chinese and Kalmegh in India, which expresses its strong bitter taste. In North India, *A. paniculata* is known as Maha-tikta, literally meaning 'king of bitters' (Benoy *et al.*, 2012; Subramanian *et al.*, 2012). Its medicinal uses have included the treatment of a variety of inflammation diseases such as diarrhea, fevers, laryngitis, gastric infections and upper respiratory tract infections, as well as other chronic and infectious disorders (Lim *et al.*, 2012). *A. paniculata* has been included in the WHO monographs on selected medicinal plants (World Health Organization, 2002), where it was reported that the plant was considered highly efficacious in inhibiting the spread of the influenza epidemic of 1919 in India (Amroyan *et al.*, 1999; Avani and Rao, 2008; Lim *et al.*, 2012; Sudhakaran, 2012; Sc, 2014). Phytochemical data show that *A. paniculata* can produce high quantities of biologically active *ent*-labdane-type diterpenoids and flavones, such as andrographolide, neoandrographolide, and andrographidine A, B and C (Koteswara Rao *et al.*, 2004; Wang *et al.*, 2009; Chao and Lin, 2010; Ma *et al.*, 2010; Subramanian *et al.*, 2012). The main bioactive compounds andrographolide and neoandrographolide are derived from the *ent*-copalol (*ent*-labda-8(17),13-dien-15-ol) diterpene backbone. This undergoes two regio-specific oxygenations, at C16 and C19, as well as heterocyclization to a 16,15-lactone, to form andrograpanin. Neoandrographolide is produced from this aglycone by glucosylation at the C19-hydroxyl, whereas andrographolide is formed by two further hydroxylations at C3 and C14. Modern pharmacological evidence suggests that *A. paniculata* as well as its diterpenoid constituents has anti-inflammatory, antimalarial, anticancer, immunomodulatory, antidiabetic and other effects (Singha *et al.*, 2003; Ajaya Kumar *et al.*, 2004; Reyes *et al.*, 2006; Sheeja *et al.*, 2006; Chandrasekaran *et al.*, 2010). This efficacy against inflammatory diseases has translated to a wide range of medical applications of *A. paniculata* in China and India.

Despite the medical importance of *A. paniculata*, in part because of the lack of a genome sequence, the pathways for biosynthesis of the bioactive diterpenoids such as andrographolide and neoandrographolide remain largely unknown. Here, we present a high-quality genome for *A. paniculata*, with an N50 (the minimum contig length needed to cover 50% of the genome) of 388 Kb, assembled using Pacbio single-molecule, real-time (SMRT) long reads. Hi-C technology was then used to anchor more than 95% of this sequence to 24 pseudo-chromosomes. Based on the

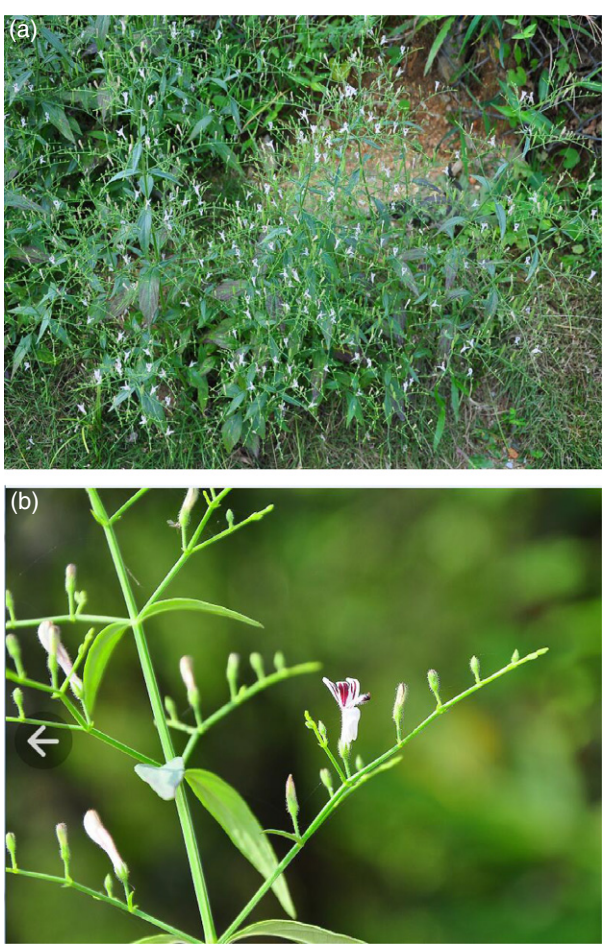


Figure 1. Characteristics of the *Andrographis paniculata* plant.
(a) Plant growth habit.
(b) Close-ups of leaf and flower. [Colour figure can be viewed at wileyonline library.com].

© 2018 The Authors

mining of this genome sequence, as well as transcriptomic sequence data generated from different tissues and from seedlings treated with methyl jasmonate (MeJA), we describe the *in vitro* functional characterization of four members of the *A. paniculata* diterpene synthase family and uncover the encoded diversity of labdane-related diterpene backbones. Additionally, we biochemically screened nine glycosyltransferases and found one that also catalyzes the conversion of andrograpanin to neoandrographide *in vitro*.

RESULTS AND DISCUSSION

Genome sequencing and assembly

Both flow cytometry and *k*-mer analysis were used to estimate the size of the *A. paniculata* genome prior to whole-genome sequencing. Flow cytometry gave an estimated haploid genome size of 280 Mb, whereas *k*-mer analysis yielded a very similar estimate of 281.26 Mb (Figures S1 and S2). We used Pacbio SMRT long-read sequencing to assemble the *A. paniculata* genome (Table S1). The size of the assembled genome was approximately 269 Mb, with 1278 contigs (longest of 2.07 Mb) and with an N50 of 388 Kb (Table 1). Nearly all contigs have a length of more than 10 Kb, and 23 contigs (1.8%) have a length over 1 Mb. The N50 of this Pacbio-only assembly is comparable or even better than recently finished plant genomes assembled with both Illumina and Pacbio reads (Table S2). We further generated approximately 270 million Illumina paired-end reads (40.2 Gb), more than 96% of which could be mapped to this *A. paniculata* draft genome. In addition, we generated transcriptomes assembled from Illumina RNA-seq reads and mapped them to the draft genome. Of the 40 046 transcripts of >1000 bp, 35 413 (88.43%) could be mapped to a single contig with >90% sequence coverage, and 38 376 (95.83%) could be mapped with >50% sequence coverage (Table S3). Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was also used to estimate the completeness of the genome assembly, and 89.4% (271/303) of eukaryote gene sets were classified as complete (Table S3). These results indicated the successful assembly of the major genic regions of the *A. paniculata* genome. Hi-C technology was then used to anchor the contigs to pseudomolecules (Figure S3). More than 95% (257 Mb) of the assembled contigs could be anchored to 24 pseudochromosomes (Table S4), with maximum and minimum lengths of 14.7 and 7.7 Mb, respectively.

Table 1 Summary of genome assembly

Number of contigs	1278
Total contigs length	269 306 334
Longest contig	2 074 285
Shortest contig	4221
Contigs > 10 Knt	1271 (99.45%)
Contigs > 100 Knt	737 (57.67%)
Contigs > 1 Mnt	23 (1.80%)
N50	388 864
L50	208
N80	165 327
L80	522
GC content	0.3332

N50 indicates the sequence length of the shortest contig at 50% of the total genome length. L50 count is defined as the smallest number of contigs whose length sum makes up half of genome size. N80 is defined as the sequence length of the shortest contig at 80% of the total genome length. L80 count is defined as the smallest number of contigs whose length sum makes up 80% of genome size. GC content: GC-content is the percentage of nitrogenous bases that are either guanine or cytosine of the genome.

Andrographis paniculata genome annotation and evolution

The *A. paniculata* genome was annotated by a combination of ab initio, homology-based and RNA-Seq assembly-based analyses. This led to the prediction of 25 428 protein-coding genes, with an average exon number of 5.79, transcript length of 3412 bp and CDS length of 1255 bp (Table S5). We then compared these genes with that of related species (Figure 2a, b). The number of gene families in *A. paniculata* (12 072) is similar to that of other plants: e.g. *Sesamum indicum* (13 247), *Arabidopsis thaliana* (13 076), and *Uricularia gibba* (11 513) (Table S6). We found a total of 587 families were specific to *A. paniculata* when compared with the genomes of *Vitis vinifera*, *Solanum lycopersicum*, and *Solanum tuberosum* (Figure 2a; Table S7). The divergence of *A. paniculata* and 11 other plants, based on their genomes, was estimated from 305 one-to-one orthologous genes (Figure 2c), with the split between *A. paniculata* and the ancestor of *Mimulus guttatus* and *Sesamum indicum* estimated to be approximately 58.8 Mya. Gene gain and loss was analyzed for *A. paniculata* relative to these 11 other species. In the *A. paniculata* lineage, 5383 gene families appear to be undergoing contraction, whereas one-quarter of this number (1290) appear to be expanding (Figure 2d). Genes in these expanded families were annotated by InterProScan, followed by a functional enrichment analysis of these genes (Table S8). Interestingly, among these were the terpene synthase (TPS) (IPR001906, $P = 0.00013$) and cytochrome P450 (CYP) monooxygenase (IPR001128, $P = 0.00043$) families, which are important in secondary metabolism, such as that leading to andrographolide and neoandrographolide.

Repetitive sequences were identified using a combination of *ab initio* and homology-based approaches. There appear to be fewer repeat elements in *A. paniculata* (53.3% of assembly) than in *S. lycopersicum* (63.2%) and *S. tuberosum* (54.5%), but more than in *Sesamum indicum* (28.5%) and *V. vinifera* (41.4%). The two most numerous types of long terminal repeat (LTR) retrotransposons in

© 2018 The Authors

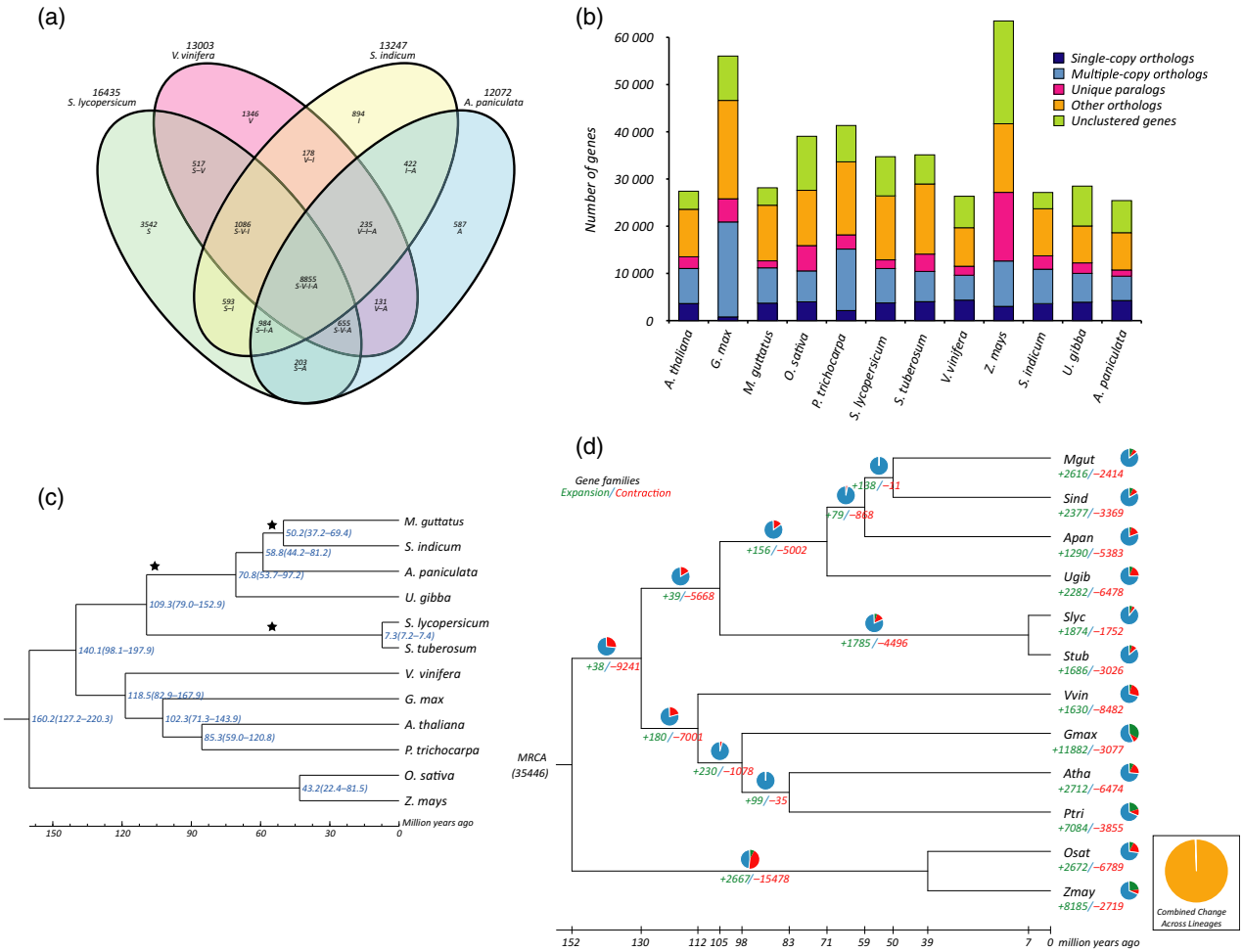


Figure 2. Evolution of the *Andrographis paniculata* genome.
(a) Venn diagram of shared orthologous gene families among four species, i.e. *Solanum ycopersicum*, *Vitis vinifera*, *Sesamum indicum* and *Andrographis paniculata.*
(b) Proportion of orthologous genes in 12 plant genomes.
(c) Estimation of the time of divergence (with error range shown in parentheses) of *A. paniculata* and 11 other species based on orthologous single-copy gene pairs. Stars highlight the location of whole-genome duplication (WGD) events.
(d) Expansion and contraction of gene families among 12 plant genomes. The pie diagram on each branch and node corresponds to combined changes across lineages. Mgut, *Mimulus guttatus*; Sind, *Sesamum indicum*; Apan, *Andrographis paniculata*; Ugib, *Uricularia gibba*; Slyc, *Solanum lycopersicum*; Vvin, *Vitis vinfera*; Gmax, *Glycine max*; Atha, *Arabidopsis thaliana*; Ptri, *Populus trichocarpa*; Osat, *Oryza sativa*; Zmay, *Zea mays*. [Colour figure can be viewed at wileyonline library.com].

A. paniculata are Ty3/Gypsy-like LTRs (10.54%) and Ty3/Copia-like LTRs (8.42%) (Table S9). Synonymous substitution rates (K_s) among paralogs and orthologs were calculated to explore possible paleopolyploidy in the *A. paniculata* genome. This analysis indicates that *A. paniculata*, *S. lycopersicum*, *Sesamum indicum* and *V. vinifera* share one paleopolyploid event, referred to as the paleohexaploidy (γ) event common to all eudicots (Figure S4) (Tang *et al.*, 2008; Vekemans *et al.*, 2012). After the split of *A. paniculata* and *S. indicum* (~59 Mya; Figure 2c), *Sesamum indicum* and *S. lycopersicum* each experienced a lineage-specific whole-genome duplication (WGD) (Tomato Genome Consortium, 2012; Wang *et al.*, 2014), but no clear peak could be seen in *A. paniculata* (Figure S4). These results jointly imply that *A. paniculata* did not undergo WGD after its split with *Sesamum indicum*.

Andrographis paniculata diterpenoid profiles

To correlate *A. paniculata* diterpenoid metabolism with gene expression, transcriptome (RNA-Seq) analysis was carried out for seedlings treated with MeJA. This mimics the wounding response and induces such secondary metabolism. In addition, the content of nine diterpenoid lactones was also targeted for analysis in these induced seedlings by ultra-high-performance liquid chromatography coupled with triple-quadrupole tandem mass spectrometry (UHPLC/MS/MS) (Table 2). Among these, 14-deoxyandrographolide, andropanolide and

© 2018 The Authors

andrographolide were the predominant metabolites found. The aglycone of neoandrographolide, andrograpanin, was only found in trace levels compared with neoandrographolide, indicating highly efficient glucosylation. All detected diterpenoids exhibited increases relative to control plants, indicating that MeJA induces their production (Jian and Wu, 2005; Chen *et al.*, 2006, 2007; Kim *et al.*, 2006; Wang *et al.*, 2010).

Diterpenoid backbone biosynthesis

Terpenoids are derived from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) precursors synthesized in plants by the mevalonate (MVA) pathway in the cytosol, or by the 2-*C*-methyl-D-erythritol-4-phosphate (MEP) pathway in the plastids (Vranová *et al.*, 2013). The formation of the general diterpenoid precursor (*E*,*E*,*E*)-geranylgeranyl diphosphate (GGPP) involves via coupling of IPP to DMAPP by the action of GGPP synthases (GGPPS) (Ma *et al.*, 2012). From the genome assembly, 15 genes were predicted to be involved in the MVA pathway and 14 genes were predicted to be involved in the MEP pathway, along with 13 GGPPS-like genes, all found using the KEGG database (Kanehisa and Goto, 2000) (Table S10). Notably, at least one gene for each enzyme in the MEP and MVA pathways showed increased transcript levels following MeJA treatment (Table S11). The plastidial MEP pathway supplies the precursors for plant diterpene biosynthesis (Lange *et al.*, 2000). Among the relevant genes identified here from *A. paniculata* are those encoding five DXS (1-deoxy-D-xylulose 5-phosphate synthases), two HDR (4-hydroxy-3-methylbut-2-enyl diphosphate reductase), two HDS (4-hydroxy-3-methylbut-2-enyl diphosphate synthase) and two DXR (1-deoxy-D-xylulose 5-phosphate reductoisomerase), with one each of MCT (2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase), CMK [4-(cytidine 5′ diphospho)-2-*C*-methyl-D-erythritol kinase] and MDS (2-*C*-methyl-D-erythritol 2,4-cyclodiphosphate synthase), indicating that genes encoding all seven MEP pathway enzymes have been identified. Protein subcellular localization prediction of these MEP pathway genes indicates that the encoded enzymes are most likely located in chloroplasts, further supporting their role in diterpenoid biosynthesis (Table S11). In addition, although transcripts for most of the GGPPS-like genes were not significantly expressed in the tissues analyzed, four did appear to be upregulated after treatment with MeJA ($P < 0.001$) (Table S11). These results suggest that the encoded enzymes might be involved in the production of andrographolide and related diterpenoids.

The bioactive diterpenoid lactones from *A. paniculata* fall into the labdane-related superfamily, the biosynthesis of which is characterized by initial bicyclization of GGPP by a class-II diterpene cyclase (Zi *et al.*, 2014a). These most often produce copalyl/labdadienyl diphosphate (CPP), and are then termed CPP synthases (CPSs) (Figure 3). The resulting bicycles generally undergo further cyclization and/or rearrangement catalyzed by class I diterpene synthases, which are termed kaurene synthase-like (KSL) based on their homology to the *ent*-kaurene synthases found in all vascular plants for gibberellin phytohormone biosynthesis (Zi *et al.*, 2014a) (Figure 3). Despite their distinct catalytic activity, CPSs and KSLs are phylogenetically related, falling into the larger TPS family, reflecting an ancient gene fusion event, and form the TPS-c and TPS-e subfamilies, respectively (Chen *et al.*, 2011). To investigate the labdane-related diterpenoid biosynthetic capacity of *A. paniculata*, we mined its genome for CPSs and KSLs, much as previously reported for *Salvia miltiorrhiza* (Ma *et al.*, 2012; Xu *et al.*, 2016). In total, five putative CPSs and two putative KSLs were found. Of the CPSs (ApCPS1-5) only three (ApCPS1-3) appeared to be full length and contained the characteristic DxDD aspartate-rich motif, whereas both KSLs (ApKSL1 and ApKSL2) appeared to be complete and contained the characteristic DDxxD motifs (Figures S5 and S6; Tables S10 and S11). Phylogenetic analysis of the translated protein with other previously functionally characterized CPSs and KSLs showed the expected grouping of CPS1–CPS4 and ApKSL1 and ApKSL2 with class-II and class-I diterpene cyclases/synthases, respectively (Figure 4a; Table S12).

Table 2 The contents of diterpenoids in *Andrographis paniculata* seedlings after treatment with MeJA

	0 h	24 h	48 h
Main compounds (mg g^{-1})			
Andrographolide	0.326 (0.095)	0.779 (0.305)	0.868 (0.252)
Andropanolide	0.959 (0.54)	1.532 (0.689)	2.328 (0.681)
14-Deoxyandrographolide	1.990 (0.133)	3.443 (0.358)	3.117 (0.203)
Minor compounds (μg g^{-1})			
Neoandrographolide	7.785 (0.54)	37.402 (4.72)	36.115 (2.484)
Andrograpanin	0.277 (0.092)	0.413 (0.286)	0.531 (0.113)
Andropanoside	27.327 (0.759)	72.775 (14.096)	71.067 (7.793)
14-Deoxy-11-hydroxyandrographolide	19.536 (1.417)	28.067 (6.216)	20.297 (1.103)
14-Deoxy-17-hydroxyandrographolide	3.629 (0.066)	5.569 (1.454)	4.69 (1.114)
14-Deoxy-11,12-dideroxyandrographoside	0.024 (0.005)	0.033 (0.002)	0.034 (0.0206)

© 2018 The Authors
The Plant Journal © 2018 John Wiley & Sons Ltd,

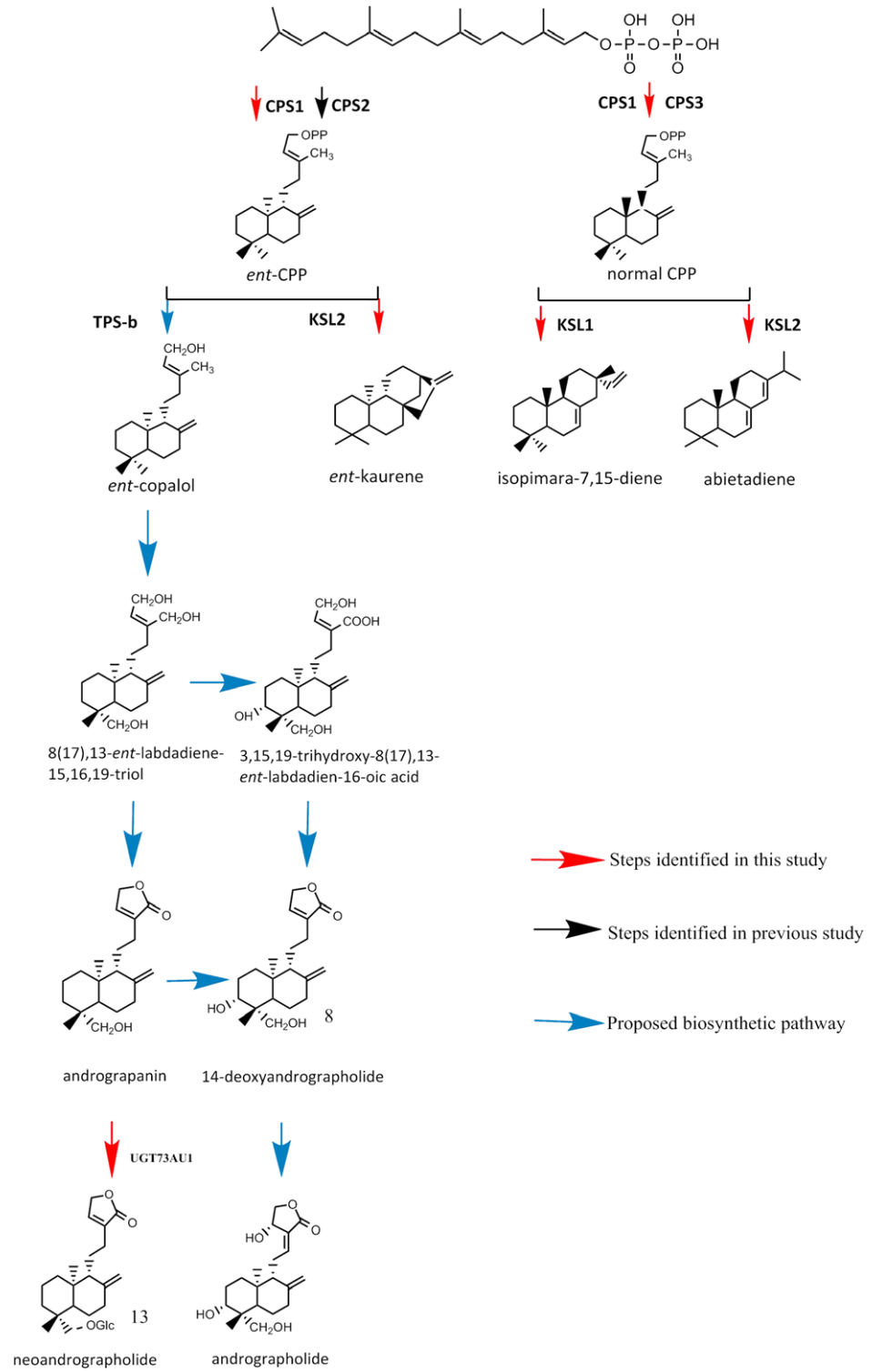


Figure 3. A general proposed pathway for diversified diterpenoid scaffolds and the major constituents andrographolide and neoandrographolide. Five diterpene synthases including CPS1– CPS3 (CPS2 previously characterized) and KSL1 and KSL2 form distinct diterpene scaffolds. A proposed biosynthesis of andrographolide, neoandrographolide and other labdane diterpenoids based on known compounds from *Andrographis paniculata* is represented. *ent*-Copalol intermediates are further modified by cytochromes P450 and other enzyme classes to afford the array of specialized diterpenoids. [Colour figure can be viewed at wileyonlinelibrary.com].

© 2018 The Authors

ApCPS2 has been previously characterized to function as an *ent*-CPP synthase (Misra *et al.*, 2015). The full-length CPSs newly identified here, ApCPS1 and ApCPS3, were functionally characterized by incorporation into a previously described modular metabolic engineering system (Cyr *et al.*, 2007). This enabled analysis by expression in *Escherichia coli* that was also engineered to produce GGPP. Both ApCPS1 and ApCPS3 were found to produce CPP, of either *ent*- or normal stereochemistry, as indicated by the detection of the dephosphorylated derivative copalol from these cultures by GC-MS (Figure 4b). To ascertain the absolute configuration of their products, these were co-expressed with class-I diterpene synthases specific for either normal CPP (i.e. a mutant of the bifunctional abietadiene synthase from *Abies grandis* that cannot produce CPP, simply termed AS hereafter; Peters *et al.*, 2001) or for *ent*-CPP (i.e. the *ent*-kaurene synthase from *Arabidopsis thaliana*, simply termed KS hereafter; Yamaguchi *et al.*, 1998). The resulting products were then compared with those generated by CPSs of known stereospecificity, as previously described (Cyr *et al.*, 2007). Notably, ApCPS1 produces both *ent*- and normal CPP, as demonstrated by its ability to supply substrate to both AS and KS (Figure 4c). This represents the first native class-II diterpene cyclase with such clearly mixed stereochemical product outcome. By contrast, ApCPS3 is stereospecific, only producing CPP of normal stereochemistry (Figure 4d), as is more typically found with these enzymes (Peters, 2010; Zi *et al.*, 2014b).

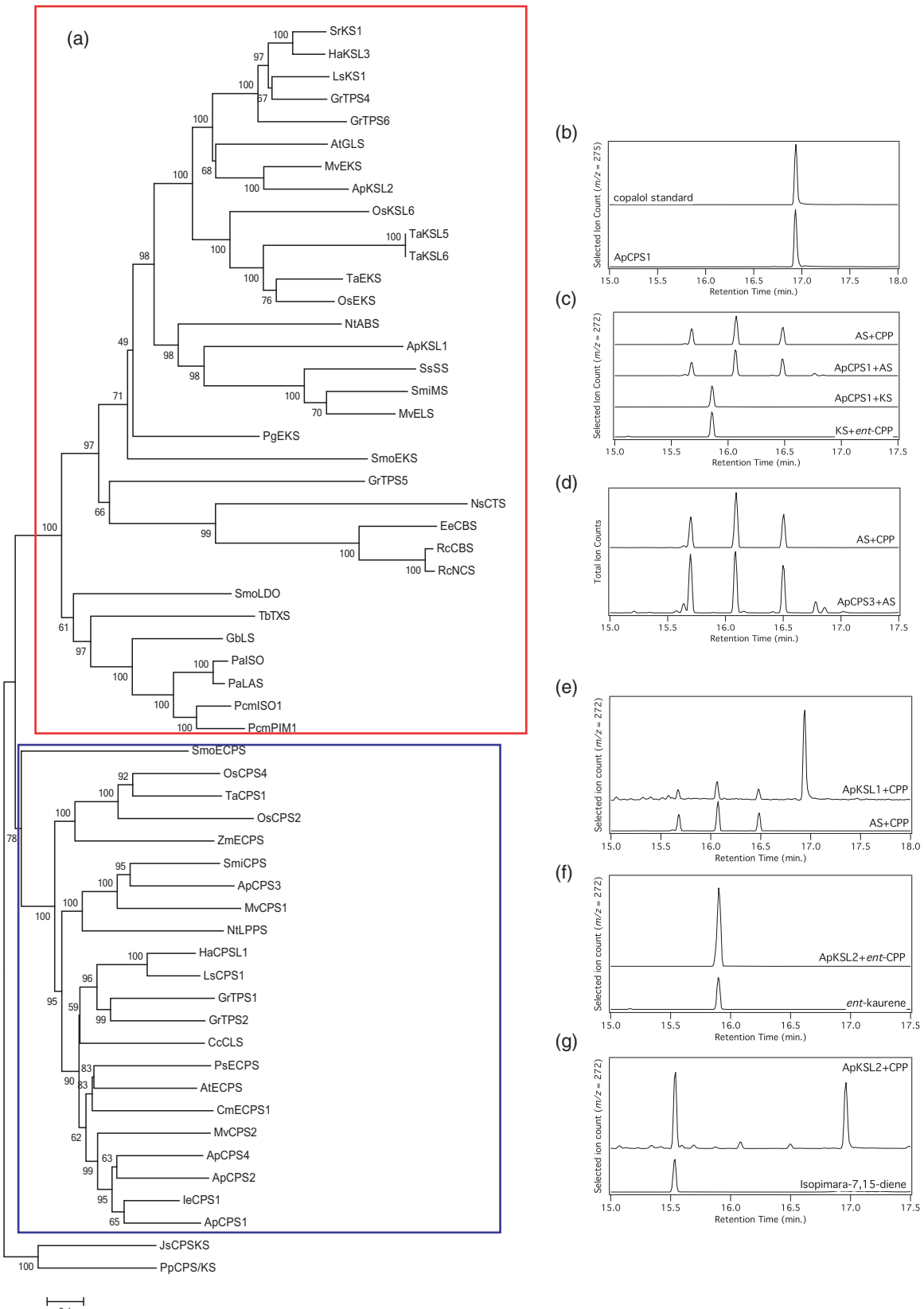
Although the characteristic DxDD motif acts as the catalytic acid (Prisic *et al.*, 2007), previous work with class-II diterpene cyclases has highlighted the importance of the pair of residues that act as the catalytic base in determining product outcome. In particular, it has been noted that the CPSs that produce *ent*-CPP for the production of the gibberellin phytohormones contain a conserved histidine and asparagine that form this catalytic dyad (Potter *et al.*, 2014). These two residues are conserved in both ApCPS1 and ApCPS2. This is consistent with the previously reported selective production of *ent*-CPP by ApCPS2 (Misra *et al.*, 2015), but is perhaps somewhat surprising for ApCPS1 given the mixture of *ent*- and normal CPP it was shown to produce here. Neither of these residues is conserved in ApCPS3, which selectively produces normal CPP. This is consistent with previous studies that have found an alternative pair of residues acting as the catalytic base, at least in normal CPP producing bifunctional diterpene synthases from gymnosperms (Criswell *et al.*, 2012; Mafu *et al.*, 2015).

The full-length KSLs identified here, ApKSL1 and ApKSL2, were similarly functionally characterized by incorporation in the same modular metabolic engineering system. In this case, by co-expression in *E. coli* also engineered to produce normal, *ent*- or *syn*-CPP, as previously described (Cyr *et al.*, 2007). ApKSL1 was found to react specifically with normal CPP (Figure 4e), and produce the same mixture of abietadienes previously reported with AS (Peters *et al.*, 2000). On the other hand, ApKSL2 acted on both *ent*-CPP, producing *ent*-kaurene (Figure 4f), and normal CPP, producing isopimara-7,15-diene (Figure 4g). It should be noted, however, that ApKSL2 seems to react more readily with normal CPP, rather than *ent*-CPP, as indicated by the co-production of *ent*-copalol, which is derived from the ability of the endogenous dephosphatases to efficiently compete with ApKSL2 for *ent*-CPP but not normal CPP. Thus, it must be noted that neither of these ApKSLs produces the expected *ent*-copalol or *ent*-labdatriene precursor to andrographolide and related diterpenoids. Given the recently reported finding that terpene synthases from the plant TPS-b subfamily can act upon CPP (Hansen *et al.*, 2016; Inabuy *et al.*, 2017), it seems at least possible that the *ent*-copalol/labdatriene synthase may be found in either this or the other TPS subfamilies represented in the *A. paniculata* genome.

Oxidational decoration of the diterpene scaffold by *A. paniculata* cytochrome P450 and 2-oxoglutarate-dependent dioxygenase

From common backbones, diterpene scaffolds are often further modified to diterpenoids, prototypically by cytochrome P450 monooxygenases (CYPs), although 2-oxoglutarate-dependent dioxygenases (2OGDs) appear to play a role in certain cases as well (Swaminathan *et al.*, 2009; Wang *et al.*, 2012a; Kakizaki *et al.*, 2017; Xu and Song, 2017). Thus, although the biosynthesis of andrographolide and neoandrographolide is believed to rely on CYPs (Garg *et al.*, 2015; Cherukupalli *et al.*, 2016), 2OGDs may also play a role. In the *A. paniculata* genome, 278 CYP genes were found by automated matching to the relevant Hidden Markov Model (HMM), PF00067, with subsequent manual assignment to the appropriate CYP families (i.e. by Prof. David Nelson, University of Tennessee). This is similar to the number found in other angiosperms (Nelson and Werck-Reichhart, 2011). Members of the CYP71 and CYP76 families have been shown to be involved in labdane-related diterpenoid biosynthesis in other Lamiales species (Banerjee and Hamberger, 2018). Genes for a total of 43 CYP71 and 14 CYP76 family members were found in the *A. paniculata* genome (Table S14). Expression analysis indicated that transcripts for 18 CYP71 and six CYP76 family members accumulate after treatment with MeJA ($P < 0.05$) (Table S13), suggesting that some of these may play a role in the biosynthesis of andrographolide and related diterpenoids. A total of 112 putative 2OGDs were found in the *A. paniculata* genome by automated matching to the 2OG-FeII_Oxy DTT domain (PF03171). This is similar to the number of 2OGDs found in other plants, such as *Oryza sativa* (114) and *Arabidopsis thaliana* (130). The

© 2018 The Authors
The Plant Journal © 2018 John Wiley & Sons Ltd,



© 2018 The Authors

2OGDs have been phylogenetically divided into three functionally distinct classes (Xu and Song, 2017), and *A. paniculata* has members of all three: i.e. DOXA (4), DOXB (21) and DOXC (90) (Table S14). DOXC class members were further clustered into 12 families, named based on known function: i.e. AOP (6), GA20ox (5), GA3ox (1), GA2ox (10), DAO (4), F3H (3), FLS/ANS (3), CODM/NCS (16), ACO (9), D4H/GSLOH/BX6 (18), H6H (3), S3H (5) and unknown (7) (Figure S7). Expression analysis indicated that transcripts for 17 2OGDs accumulated after treatment with MeJA ($P < 0.05$), suggesting that these might also be involved in the biosynthesis of andrographolide and related diterpenoids (Table S14).

Putative diterpenoid biosynthestic gene clusters

With the rapidly growing number of available plant genomes, it has become evident that these often contain gene clusters that encode, at least partially, biosynthetic pathways for specialized metabolites (Nuetzmann and Osbourn, 2015). To determine whether *A. paniculata* contains any such clusters for diterpenoid biosynthesis, the genome was assessed as to whether other biosynthetic genes are found near those for the characterized ApCPSs or ApKSLs. Intriguingly, adjacent to *ApCPS3* are two encoding CYP71 family members, whereas on a separate contig three genes encoding CYP714 family members are linked to *ApKSL1* (Figure S8). In addition, *ApCPS4* and *ApCPS5* are found together with genes encoding four CYP76 family members on yet another contig. It can then be hypothesized that these CYPs may be involved in the further elaboration of the initial hydrocarbon scaffolds produced by the nearby (co-clustered) CPS(s) or KSL.

UDP-dependent glycosyltransferases in *A. paniculata* genome

Following sequential oxidation by CYP450 and 2OGD, glycosylation by UDP-dependent glycosyltransferases (UGT) is often observed in secondary metabolism. The added glycan moiety confers pharmacological bioactivity, reduced toxicity and increased solubility (Vogt and Jones, 2000; Jones and Vogt, 2001; Lorenc-Kukuła *et al.*, 2004). One of the major bioactive constituents of *A. paniculata*, neoandrographolide, is just such a glycosylated diterpenoid lactone. In this study, we used the 44 amino acids in the PSPG motifs and full-length amino acid sequence of AtUGT71C5 (*Arabidopsis thaliana*), MtUGT73K1 (*Medicago truncatula*), SgUGT74AC1 (*Siraitia grosvenorii*), SrUGT76G1 (*Stevia rebaudiana*), SrUGT85C2 (*S. rebaudiana*) and AtUGT88A1 (*Arabidopsis thaliana*) to probe *A. paniculata* sequence information. A total of 120 putative *UGT* genes were identified, including 99 genes encoding putative full-length UGTs (i.e. that are more than 250 amino acids in length) (Table S15). Although this is less than the number of characterized *UGT* genes in *Arabidopsis thaliana* (120), *Medicago truncatula* (187), *Cypripedium arietinum* (125), *Glycine max* (242), *Gossypium hirsutum* (196) and *Lotus japonicus* (188) (Yonekura-Sakakibara and Hanada, 2011; Huang *et al.*, 2015; Yin *et al.*, 2017), it may reflect less glycoslation in this plant species or the incomplete nature of the draft genome assembly. The full-length *A. paniculata UGT* genes were named based on the accepted naming convention for signature PSPG motifs (http://prime.vetmed.wsu.edu/resources/udp-glucuronsyltransferase-homepage) (Mackenzie *et al.*, 2005). A phylogenetic tree constructed from the 120 screened *A. paniculata* UGT proteins as well as all 120 UGT proteins from *Arabidopsis thaliana* revealed that the *ApUGT* genes clustered into 22 families. Using the classification scheme employed for *Arabidopsis thaliana* (Ross *et al.*, 2001), these ApUGTs were clustered into 13 groups (A–M) (Figure S9).

The metabolic and transcriptomic analyses of the response of *A. paniculata* seedlings to MeJA was used to highlight ApUGTs that might produce neoandrographolide, which is glucosylated andrograpinin. Similar to the accumulation of the other diterpenoid lactones, neoandrographolide content was significantly (4.8-fold) increased within 24 h compared with controls (Figure 5a). Based on previous reports of UGTs involved in terpenoid

Figure 4. Functional and stereochemical analysis of *ApCPS*s and *ApKSL*s.
(a) Phylogenetic analysis of the CPS and KSL genes in different plants. Neighbor-joining trees of class-II and class-I diterpene synthases based on aligned protein sequences. The phylogenetic tree was rooted with JsCPSKS and PpCPS/KS. Protein sequences are shown in Table S12.
(b) Production of CPP (of either *ent*- and/or normal stereochemistry). GC-MS chromatograms of extracts from cultures expressing ApCPS1 or ApCPS3 (as indicated) in *Escherichia coli* also engineered to produced GGPP. CPP is detected as the dephosphorylated copalol, presumably generated by endogenous phosphatases, with verification by comparison of retention time and mass spectra to an extract from a strain analogously engineered with a known CPS.
(c) ApCPS1 produces both *ent*- and normal CPP. GC-MS chromatograms of extracts from cultures in which ApCPS1 is expressed in *E. coli* also engineered to produce GGPP and additionally co-expressing either the normal CPP-specific AS or *ent*-CPP-specific KS, in comparison with CPSs of known stereospecificity.
(d) ApCPS3 produces normal CPP. GC-MS chromatograms of extracts from cultures in which ApCPS1 is expressed in *E. coli* also engineered to produce GGPP and additionally co-expressing the normal CPP specific AS, again with verification by comparison with an extract from an analogously engineered strain with CPS known to produce normal CPP.
(e) ApKSL1 is specific for normal CPP and produces the same mixture of abietadienes as AS. GC-MS chromatograms of extracts from cultures expressing ApKSL1 or AS in *E. coli* also engineered to produce CPP.
(f) ApKSL2 readily reacts with *ent*-CPP. GC-MS chromatograms of extracts from cultures expressing ApKSL1 or KS in *E. coli* also engineered to produce *ent*-CPP.
(g) ApKSL2 also reacts with normal CPP. GC-MS chromatograms of extracts from cultures expressing ApKSL1 or AS in *E. coli* also engineered to produce CPP. [Colour figure can be viewed at wileyonlinelibrary.com].

© 2018 The Authors

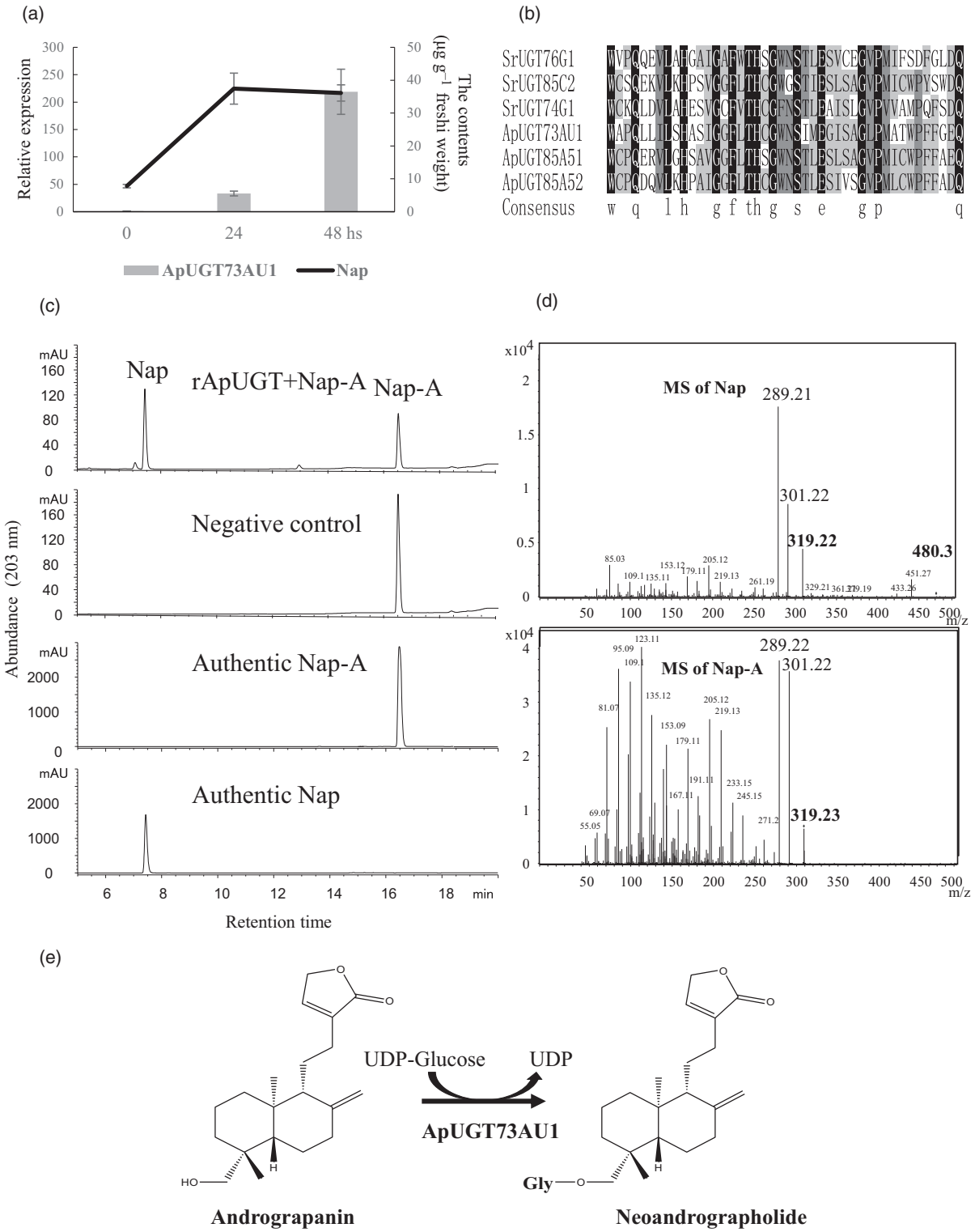


Figure 5. Characterization of rApUGT73AU1 to catalyze andrograpanin to neoandrographolide.
(a) The expression level of ApUGT73AU1 and the neoandrographolide (Nap) content in seedlings after treatment with MeJA.
(b) Multiple sequence alignment of PSPG motifs of ApUGTs and SrUGTs.
(c) Representative UPLC analysis showing the in vitro production of neoandrographolide (Nap) by incubating rApUGT73AU1 with andrograpanin (Nap-A).
(d) Representative indicate the mass spectrum of Nap (upper panel) and andrograpanin (Nap-A) for rApUGT73AU1.
(e) The proposed pathway for Nap biosynthesis based on the functions of rApUGT73AU1.

© 2018 The Authors

biosynthesis (Poppenberger *et al.*, 2005; Richman *et al.*, 2005; Sayama *et al.*, 2012), it was hypothesized that neoandrographolide would be produced by a UGT from the UGT71, UGT73, UGT76, UGT85 or UGT88 families. Accordingly, we focused on the 29 ApUGTs from these families, and found that mRNA levels were significantly elevated (more than twofold) by treatment with MeJA for a number of these. Namely, CXN00023829 (82.8-fold), CXN00009930 (7.5-fold), CXN00012854 (25.3-fold), CXN00006409 (188.5-fold), CXN00006414 (5.5-fold), CXN00009309 (7.4-fold), CXN00003963 (7.1-fold) and CXN00013979 (4.1-fold). In addition, seven genes were found to be highly expressed in leaves. Namely, CXN00004677, CXN00012856, CXN00003107, CXN00000415, CXN00000414, CXN00017140 and CXN00012855. To functionally investigate the ability of these to produce neoandrographolide, nine of these genes were successfully cloned (CXN00006414, CXN00000414, CXN00000415, CXN00003963, CXN00003107, CXN00013979, CXN00012856, CXN00017140 and CXN00004677), which were named *UGT73AU1*, *UGT76T1*, *UGT76T2*, *UGT85A51*, *UGT85A52*, *UGT85E6*, *UGT88A14*, *UGT88D10* and *UGT88L1*, respectively. These were then expressed in *E. coli*, and the recombinant proteins purified and evaluated with enzymatic assays (Figure S10). Sequence alignment of these nine ApUGTs with the UGTs from *Stevia rebaudiana* that play a role in glycosylating a similar labdane-related diterpenoid (Richman *et al.*, 2005; Brandle and Telmer, 2007) revealed a conserved domain containing the PSPG motif near their C-terminal domain (Figures 5b and S10). The last glutamine (Q) residue within this PSPG motif is thought to confer specificity for UDP-glucose as the sugar donor for glycosylation (Kubo *et al.*, 2004). Notably, three ApUGTs possess this Q, suggesting that they may all use UDP-glucose as a sugar donor. Two sugar donors (UDP-glucose and UDP-glucuronic acid) and three andrographolide-like aglycones (andrographolide, andrograpanin and 14-deoxy-11,12-didehydroandrographolide) were tested as substrates. We found that of the UGTs tested, only recombinant UGT73AU1 exhibited catalytic activity, specifically with andrograpanin and 14-deoxy-11,12-didehydroandrographolide and UDP-glucose as the sugar donor (Figures 5c and S11). No activity was detected for the remaining eight recombinant UGT proteins with any of the tested aglycone substrates, regardless of sugar donor. The enzymatic products analyzed via UPLC showed new products were produced with andrograpanin and 14-deoxy-11,12-didehydroandrographolide (Figure S11), compared with the corresponding controls. UPLC quadrupole time-of-flight mass spectrometry (Q-TOF MS) analysis revealed that these products contained fragments with an *m/z* of 162 and were characterized by the loss of one glucose. This loss yielded corresponding aglycones with each substrate (Figures 5d and S11), indicating they were andrographolide-like monoglucosides. When compared with reference standards, it was verified that these compounds were neoandrographolide and 14-deoxy-11,12-didehydroandrographoside, derived from andrograpanin and 14-deoxy-11,12-didehydroandrographolide, respectively. UGT73AU1 exhibited higher enzymatic activity with andrograpanin (1 nkat mg^{-1}) than 14-deoxy-11,12-didehydroandrographolide (0.35 nkat mg^{-1}). Given its observed activity, UGT73AU1 seems to discriminate against the 14-hydroxyl found in andrographolide, which does not serve as a substrate, as this is not found in the andrograpanin and 14-deoxy-11,12-didehydroandrographolide that it does react with.

Putative diterpenoid-associated regulatory genes after MeJA induction

A promising approach to increasing plant secondary metabolite production is the use of transcription factors (TFs) that stimulate the gene expression of enzymes that comprise the relevant biosynthetic pathways. Several types of TFs that participate in sesquiterpenes, diterpenoid, triterpenoid and monoterpenoid alkaloid biosynthesis have been characterized, belonging to the families of AP2/ERF (APETALA2/ethylene-response factors), bHLH (basic helix-loop-helix), and WRKY (Rushton *et al.*, 2010; Yamada *et al.*, 2011; Zhang *et al.*, 2011; Yu *et al.*, 2012; Lu *et al.*, 2013; Schweizer *et al.*, 2013; Schluttenhofer and Yuan, 2015; Van Moerkercke *et al.*, 2015; Yamamura *et al.*, 2015; Shen *et al.*, 2016; Zhou *et al.*, 2016) A total of 1489 genes in *A. paniculata* are predicted to encode TFs, including 115 members of the bHLH and 67 members of the WRKY families (Table S16). Here we found that labdane-diterpenoid biosynthesis from *A. paniculata* responds strongly to treatement with MeJA, which is known to induce TFs involved in the regulation of a variety of secondary metabolite biosynthesis (van der Fits and Memelink, 2000; Van der Fits and Memelink, 2001; Zhang *et al.*, 2011; De Geyter *et al.*, 2012). To identify putative regulators of the andrographolide pathway, 94 candidate transcriptional factors were selected by mining the RNA-seq data for TFs with transcript levels that were significantly increased following treatement with MeJA ($P < 0.05$; Table S15), indicating a potential role in the regulation of such labdane-diterpenoid biosynthesis. Functional characterization of these TFs using a transient system such as virus-induced silencing (VIGS) and overexpression, in combination with metabolite analysis, will be employed in the future.

In summary, the extensive use of the whole plant of *A. paniculata* as an anti-inflammatory drug for the treatment of fever, cold, laryngitis, diarrhea and inflammation in India and China has led to a broad range of phytochemical/pharmacological studies that have reported the discovery of anti-inflammatory activity for a diverse array of diterpenoid lactones, particularly andrographolide and neoandrographolide. With the advent of next-generation

© 2018 The Authors
The Plant Journal © 2018 John Wiley & Sons Ltd,

sequencing techniques, -omics analyses of non-model plants are now feasible. Here, the *A. paniculata* genome is reported, providing a foundation for cultivar breeding. Along with the tissue-specific and MeJA-induced transcriptomes reported here, this further provides a rich data set for mining candidate genes involved in the production and regulation of *A. paniculata* bioactive diterpenoid lactone natural products. For example, the CYP and 2OGD oxygenase superfamilies in *A. paniculata* were annotated, generating a set of tailoring enzymes for the future investigation of diterpenoid lactone biosynthesis. In addition, the identification of putative transcription factors on a genome-wide scale is expected to facilitate a better understanding of the underlying regulation of such metabolism. The utility of the generated sequence data was more directly demonstrated by the functional characterization of *A. paniculata* diterpene synthases, revealing capacity for the production of a variety of labdane-related diterpene backbones. Moreover, a UGT was found that catalyzes glucosylation of andrograpanin to the bioactive neoandrographolide. Accordingly, the data reported here are expected to provide the basis for further insights into the production of bioactive diterpenoid lactone constituents of *A. paniculata*.

EXPERIMENTAL PROCEDURES

Plant materials for sequencing

Andrographis paniculata plants were cultivated in a glasshouse at the Institute of Chinese Materia Medica, China. Genomic DNA was extracted from the leaves of one *A. paniculata* plant, using the plant DNA extraction protocol (Tiangen co. Ltd., http://www.tiangen.com). For RNA-seq, RNA samples were prepared from root, vegetative stem, young leaf, open flower, mature fruit, and cotyledons of seedlings sprayed with 50 mM MeJA for 0, 24 and 48 h using RNeasy Plant Mini Kit (Qiagen, https://www.qiagen.com). The quality and quantity of the isolated DNA and RNA were checked by electrophoresis on a 0.8% agarose gel and Epoch (BioTek Instruments, https://www.biotek.com).

Estimation of the genome size

Both flow cytometry and *k*-mer analysis were used to estimate the size of the *A. paniculata* genome prior to whole-genome sequencing. Flow cytometry analysis for the measurement of nuclear DNA content was performed using a Partec CA II (Partec, Munster, Germany).

Genome sequencing and assembly

PacBio SMRTbell libraries (20-kb inserts) were prepared with the standard PacBio library preparation protocols, and the sequencing was conducted on a PacBio RS II (PacBio, https://www.pacb.com) system using P6-C4 chemistry (Novogene, https://en.novogene.com). This generated 1.2 million SMRT long reads with a total length of 11.8 Gb. The longest read has a length of 78 939 bp, whereas the mean length of all reads is 9326 bp and the median length is 9024 bp. More than 93% of the reads have lengths longer than 1 Kb, and more than 44% have lengths longer than 10 Kb (Table S3). We used CANU 1.5 (Koren *et al.*, 2017) to assemble these SMRT long reads with the following options: 'genomeSize = 280 m batMemory = 256 g'.

Hi-C assistant contig clustering

The Hi-C library was prepared by Annoroad Genomics (http://en.annoroad.com), following the standard procedure (Lieberman-Aiden *et al.*, 2009). The Hi-C sequencing data were aligned to the assembled contigs by BWA-MEM, and then the contigs were clustered onto chromosomes with LACHESIS (http://shendurelab.github.io/LACHESIS/).

Gene prediction and annotation

We employed three methods to predict the *A. paniculata* genes: (i) a homology-based method; (ii) a *de novo* method; and (iii) an expressed sequence tag (EST)/transcript-based method. MAKER was used to predict genes in the assembled genome for the homology-based approach (Cantarel *et al.*, 2008). Prior to *de novo* gene prediction, DNA- and protein-related repeat elements were masked using REPEATMAKER with Repbase (Cantarel *et al.*, 2008) and RepeatRunner databases (Tarailo-Graovac and Chen, 2009), respectively. Gene prediction and annotation by *ab initio* gene prediction used RNA- and protein-based alignments with AUGUSTUS (http://www.yandell-lab.org/software/repeatrunner.html). To validate and complete the gene predictions, transcriptomes from root, shoot, leaf, flower and fruit material were assembled using TRINITY (Haas *et al.*, 2013), with default parameters, which were aligned with the assembled genome using BLASTN (Camacho *et al.*, 2009). For the annotation of protein-coding genes, the nucleotide sequences of high confidence genes were searched against NCBI, KEGG, Pfam (Bateman, 2004) and Swissport databases, with a minimal *e*-value of $1e^{-5}$. BUSCO was then used to assess the completeness of the final genome assembly (Waterhouse *et al.*, 2018). The localizations of deduced proteins were predicted using the TargetP 1.1 server (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson *et al.*, 2007).

TEs and repetitive DNA

To predict the TEs in the *A. paniculata* genome, we employed a combined methodology, which incorporated *de novo* and homology-based methods. First, a TE library was constructed using REPEATMODELER (Tarailo-Graovac and Chen, 2009), and then REPEATMASKER was used to align the assembled genome to perform *de novo* prediction and find known TEs using a TE library composed from the Repbase database.

Phylogenetic tree and evolutionary analysis

Phylogenetic construction from the genomes of *A. paniculata* with *Arabidopsis thaliana*, *Glycine max*, *Mimulus guttatus*, *Oryza sativa*, *Populus trichocarpa*, *Sesamum indicum*, *Solanum lycopersicum*, *Solanum tuberosum*, *Utricularia gibba*, *Vitis vinifera* and *Zea mays* was carried out using MRBAYES 3.1.2 (Ronquist and Huelsenbeck, 2003), based on the HKY85 model (Hasegawa *et al.*, 1985) selected by MODELTEST 3.7 (Posada and Crandall, 1998), in which Akaike information criterion (AIC) was used to select the best model. The robustness of the retrieved tree topology was also assessed by running the maximum-likelihood method implemented in RAXML 8.2.10 (Stamatakis, 2014) (Figure S12). The divergence time of selected species was analyzed using the program MCMCTREE of PAML 4.4 (Yang, 2007). Fossil divergence time points for the *Solanum lycopersicum/Solanum tuberosum* split (7.2–7.4 Mya) (Sato *et al.*, 2012), the *Oryza sativa/Zea mays* split (>20 Mya) (Paterson *et al.*, 2004, 2009) and the *Oryza sativa/Vitis*

© 2018 The Authors

vinifera split (130–240 Mya) (Jaillon *et al.*, 2007) were used to calibrate the tree. K_s analysis was conducted by firstly using MCSCANX to obtain genomic collinear blocks (Wang *et al.*, 2012b), then program YN00 from the PAML package was used to calculate K_s values of gene pairs contained in collinear blocks. In order to obtain the gene families from the 14 species selected, ORTHOMCL 1.4 was used for gene clustering, setting the main inflation value to 1.5 and using the other default parameters (Li *et al.*, 2003). Then CAFÉ 2.1 was used for contraction and expansion analysis of these gene families (De Bie *et al.*, 2006). For phylogenetic reconstruction of the diterpene synthase, 2-oxoglutarate-dependent dioxygenase and UDP-glycosyl transferase families, neighbor-joining phylogenetic analyses with 1000 bootstrap repetitions were performed using MEGA 5.02 beta (Kumar *et al.*, 2016).

Metabolite detection

To quantify the levels of nine diterpenoid lactones in *A. paniculata* seedlings after treatment with 50 mM MeJA, UHPLC/MS/MS was used. Fresh samples were powdered in liquid nitrogen and 100 mg was extracted with 15 ml of methanol in an ultrasonic bath at 25°C for 30 min. The resulting sample solutions, as well as nine standard compounds (andrographolide, andropanolide, 14-deoxyandrographolide, neoandrographolide, andrograpanin, andropanoside, 14-deoxy-11-hydroxyandrographolide, 14-deoxy-17-hydroxyandrographolide and 14-deoxy-11,12-dideroxyandrographoside) from BioBioPha Co., Ltd http://www.biobiopha.com), were filtered through a 0.22-μm membrane, and 2-μl aliquots were injected for analysis. A 1290 series UHPLC was coupled with a 6470 triple quadrupole mass spectrometer via an AJS–ESI interface (Agilent Technologies, https://www.agilent.com). Samples were separated over an Agilent Eclipse Plus C18 column (rapid resolution high definition (RRHD) 1.8 μm, 2.1 × 50 mm). Mobile phases A and B were acetonitrile and water solutions containing 0.1% formic acid, respectively. The analytes were eluted using a linear gradient program: 0–2 min, 6→25% B; 2–6 min, 25→30% B; 6–7 min, 30→80% B; 6–10 min, 80→100% B; and 10–11 min, 100% B. The flow rate was 0.30 ml min^{-1} and the column temperature was 30°C. The mass spectrometer was operated in positive-ion mode, with a sheath gas temperature of 250°C, gas flow at 11.0 L min^{-1}, and nebulizer gas at 40 psi. The capillary voltage was set at 4000 V, nozzle voltage set to 500 V, and delta electron multiplier voltage (EMV) was to 200 V. Metabolites were detected in multiple reaction monitoring (MRM) mode, where two precursor product ion MRM transitions were selected for each compound. Data were acquired and analyzed using MASSHUNTER B.07.00 to quantify all nine metabolites for which standards were available.

Isolation of five *ApCPS/KSL*s and nine *ApUGT*s from *A. paniculata*

Using plant RNA extraction protocols (Tiangen Co. Ltd, http://www.tiangen.com), total RNA was extracted from the leaf of *A. paniculata* for cDNA synthesis. First-strand cDNA was synthesized using the PrimeScript First-Strand cDNA Synthesis Kit and oligo(dT) primer (TaKaRa, http://www.takara-bio.com). Primers were designed for three *ApCPS*s, two *ApKSL*s and nine *ApUGT*s (from different UGT families), according to the *A. paniculata* genome sequence, and used to clone these full-length genes that were inserted into pEASY blunt vector (TransGen Biotech, http://www.transgenbiotech.com) for Sanger sequencing using an ABI3730 DNA Sequencer (Applied Biosystems, now ThermoFisher Scientific, https://www.thermofisher.com) (Table S17).

UGT gene expression analysis

Total RNA was reverse transcribed into cDNA using the TransScript II One-Step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech), according to the manufacturers protocol. Ten-fold diluted cDNA was used as the template for subsequent qRT-PCR analysis using TransStart Green qPCR SuperMix (TransGen Biotech) on Rotor-Gene Q MDx (Qiagen, https://www.qiagen.com), with primer sequences that can be found in Table S17. These PCR reactions were performed using the following cycling parameters: 95°C for 7 min (enzyme activation), 35 cycles of 95°C for 15 s, 60°C for 30 s and 72°C for 30 s, followed by a melting curve cycle from 60 to 90°C. The results were normalized against actin as a reference gene. The relative transcript level was calculated as the mean of three technical replicates of three biological replicates.

Functional characterization of *ApCPS/KSL* genes

The *ApCPS1* and *ApCPS3* genes were truncated to remove the N-terminal plastid targeting sequence for expression of pseudomature enzymes in *E. coli*. These constructs were generated by PCR, and first cloned in pENTR-SD-dTOPO (Invitrogen, now ThermoFisher Scientific, https://www.thermofisher.com), with the corresponding insert verified by complete gene sequencing. Each gene was then individually subcloned by directional recombination into a previously described pGG-DEST vector (Cyr *et al.*, 2007), creating pGG-DEST::ApCPS1 and three expression constructs. These were used to examine product outcome, first by expression alone, and then by coexpression with compatible pDEST14-based AS or KS constructs to examine the stereospecific product outcome. Similarly, ApKSL1 and ApKSL2 were truncated to remove the N-terminal plastid targeting sequence for expression of pseudomature enzymes in *E. coli*. These constructs were generated by PCR, and first cloned in pENTR-SD-dTOPO (Invitrogen), with the corresponding insert verified by complete gene sequencing. Each gene was then individually subcloned by directional recombination into pDEST14, creating pDEST14::ApKSL1 and two expression constructs. These were used to examine product outcome by co-expression with previously described compatible pGGeC, pGGsC and pGGnC vectors, which lead to the production of *ent-*, *syn-*, or normal CPP, respectively, as previously described (Cyr *et al.*, 2007).

Expression and purification of recombinant UGT proteins in *E. coli*

The UGTs were subcloned to construct pMAL-UGT expression vectors, which were then transformed into *E. coli* strain Novablue competent cells. To induce protein expression, 0.3 mM of isopropyl β-D-thiogalactoside (IPTG) was added when the OD_{600} value of the cell culture (grown at 37°C) reached 0.5. After 24 h of incubation with shaking at 16°C, the cells were harvested by centrifugation at 4°C and then stored at −80°C until purification. The MBP-fusion proteins were purified using protocols for the pMAL Protein Fusion and Purification System (New England BioLabs, https://www.neb.com). The recombinant UGT proteins (5–10 μg) were incubated for 1 h at 37°C in a final volume of 50 μl comprised of 10 mM dithiothreitol (DTT), 20 mM Tris-HCl (pH 7.0), 0.5 mM substrates, and 2 mM UDP-glucose or UDP-glucuronic acid. Reactions were stopped by the addition of methanol and centrifuged at 16 000 ***g*** for 10 min, followed by analysis via UPLC. To investigate the enzymatic specificity of the recombinant UGT73AU1, purified enzyme (5–10 μg) was incubated for 30 min at 37°C in 50-μl reaction mixtures comprising 10 mM DTT, 20 mM Tris-HCl (pH 7.0), 0.5 mM substrates and 2 mM UDP-glucose.

© 2018 The Authors

Reactions were terminated by adding methanol, centrifuged at 16 000 ***g*** for 10 min, and also analyzed via UPLC. The enzymatic products were further identified by Q-TOF-MS, as described below. Acquity ultra UPLC (Agilent 1290 Infinity II) consisting of an autosampler and a binary pump was used for analysis. Specifically, the compounds were separated over an EclipsePlusC18 RRHD (1.8 μm, 2.1 × 50 mm i.d.; Agilent) analytical column at a temperature of 30°C, with a sample injection volume of 2 μl. A gradient elution was achieved using two solvents: 0.1% formic acid in water (A) or acetonitrile (B), at a flow rate of 0.3 ml min^{-1}. The gradient program consisted of: 0–2 min, 74% A; 2–4 min, 68% A; 4–12 min, 62% A; 12–20 min, 55% A; 20–22 min, 95% B; 22–24 min, 95% A for balancing, with a return to initial conditions over 1 min. The detected wavelength was 203 nm. The UPLC system used was interfaced with a 6545 Q-TOF LC/MS (Agilent), equipped with an electrospray (Turbo V TM) ion source. The MS was used in positive-ion mode with the following conditions: dual AJS ESI (segment), gas temp 325°C; drying gas, 5 L min^{-1}; nebulizer, 35 psig; sheath gas temp, 350°C; and sheath gas flow, 11 L min^{-1}. Dual AJS ESI (experiment): V_{cap}, 3500 V; nozzle voltage (experiment), 500 V. MS TOF (experiment): fragmentor, 130 V; skimmer, 65 V; Oct 1 RF V_{pp}, 750 V; and collected signal *m/z*: 100–1000.

Data submission

Accession numbers: the clean reads are deposited in Sequence Read Archive (SRA) under SRP143459.

AUTHORS' CONTRIBUTIONS

W.S., C.S., R.P. and S.L.C. designed the experiments. W.S. isolated the genomic DNA and RNA from different tissues and MeJA-treated seedlings. C.S. performed the assembly of Illumina data and evolutionary analysis. L.L. performed the assembly of PacBio sequence data. W.S., M.M.X. and Q.G.Y. cloned and characterized gene function. W.S., C.X. and S.C. performed LC mass spectrometry assays. W.S., L.L., Q.G.Y. and R.J.P. wrote the manuscript with input from the other co-authors. X.L.Z., N.X. and C.H.J. were responsible for plant germplasm collection. All authors read and approved the final article.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (81703647), the US National Institutes of Health (GM076324) and the National Postdoctoral Program for Innovative Talents Fund (BX201600155).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Evaluation of the genome size of *Andrographis paniculata* by flow cytometry.

Figure S2. Evaluation of the genome size of *Andrographis paniculata* by 17-*mer* analyses.

Figure S3. Genome wide Hi-C heat map.

Figure S4. K_s distributions of syntenic paralogs and orthologs from *Andrographis paniculata*, *Sesamum indicum, Vitis vinfera*, and *Solanum lycopersicum.*

Figure S5. Four sequence comparison of the *Andrographis paniculata* copalyl diphosphate synthase (CPS) enzymatic family with two CPS genes from *Marrubium vulgare*.

Figure S6. Two-sequence comparison of the *Andrographis paniculata* kaurene synthase-like (KSL-like) enzymatic family with three previously characterized KSL enzymes from *Oryza sativa*.

Figure S7. Analysis of the phylogenetic relationships of 2OGD gene members in *Andrographis paniculata.*

Figure S8. Physical clustering of diterpenoid pathway gene in *Andrographis paniculata*.

Figure S9. Neighbor-joining (NJ) tree of the UGT family shows 13 distinct groups (A–M), based on aligned protein sequences.

Figure S10. Multiple sequence alignment of PSPG motifs of ApUGTs and three purified recombinant UGT proteins on denaturing PAGE.

Figure S11. Characterization of rApUGT73AU1 activity with 14-deoxy-11,12-didehydroandrographolide.

Figure S12. Phylogeny of *Andrographis paniculata* and 11 other species estimated by the maximum-likelihood method.

Table S1. Sequencing summary of PacBio SMRT library.

Table S2. Summary of some plant genomes that have been published recently.

Table S3. Estimation of *Andrographis paniculata* genome by RNA-seq and BUSCO results.

Table S4. Contigs anchored to pseudo-molecules.

Table S5. Summary of gene annotation.

Table S6. Summary of gene families.

Table S7. Functional enrichment analysis of the genes specific to *Andrographis paniculata* plant using InterPro database.

Table S8. Functional enrichment analysis of the genes belonging to families specifically expanded in the *Andrographis paniculata* genome.

Table S9. Transposable element categories and contents in *Andrographis paniculata* genome.

Table S10. MEP, MVA, IPPI and GGPPS-like genes in the *Andrographis paniculata* genome.

Table S11. FPKM expression levels of MEP, MVA, IPPI and GGPPS-like genes in the *Andrographis paniculata* genome.

Table S12. The CPS and KSL protein sequences from *Andrographis paniculata* and other species for phylogenetic reconstruction in this study.

Table S13. FPKM expression levels of P450 genes from CYP71 and 76 families.

Table S14. 2OGD genes in the *Andrographis paniculata* genome and their expression pattern.

Table S15. UGT genes expression in different tissues and seedlings treated with MeJA.

Table S16. List of transcription factor genes predicted in the *Andrographis paniculata.*

Table S17. Characterization of ApCPS, KSL and UGT. This spreadsheet lists the gene names, their corresponding IDs and GenBank numbers, as well as the primers used for the cloning of every individual sequence characterized in the paper.

REFERENCES

Ajaya Kumar, R., Sridevi, K., Vijaya Kumar, N., Nanduri, S. and Rajagopal, S. (2004) Anticancer and immunostimulatory compounds from Andrographis paniculata. *J. Ethnopharmacol.* **92**, 291–295.

Amroyan, E., Gabrielian, E., Panossian, A., Wikman, G. and Wagner, H. (1999) Inhibitory effect of andrographolide from Andrographis paniculata on PAF-induced platelet aggregation. *Phytomedicine*, **6**, 27–31.

Anju, D., Jugnu, G., Kavita, S., Arun, N. and Sandeep, D. (2012) A review on medicinal prospectives of Andrographis paniculata Nees. *J. Pharm. Sci. Innov.* **1**, 1–4.

© 2018 The Authors

Avani, G. and Rao, M.V. (2008) In vitro cytogenetic effects of Andrographis paniculata (kalmegh) on arsenic. *Phytomedicine*, **15**, 221–225.
Banerjee, A. and Hamberger, B. (2018) P450s controlling metabolic bifurcations in plant terpene specialized metabolism. *Phytochem. Rev.* **17**, 81–111.
Bateman, A. (2004) The Pfam protein families database. *Nucleic Acids Res.* **32**(suppl 1), D138–D141.
Benoy, G.K., Animesh, D.K., Aninda, M., Priyanka, D.K. and Sandip, H. (2012) An overview on andrographis paniculata (burm. F.) Nees. *Int. J. Res. Ayurveda Pharm.* **3**, 752–760.
Brandle, J.E. and Telmer, P.G. (2007) Steviol glycoside biosynthesis. *Phytochemistry*, **68**, 1855–1863.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S. and Yandell, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196.
Chandrasekaran, C.V., Gupta, A. and Agarwal, A. (2010) Effect of an extract of Andrographis paniculata leaves on inflammatory and allergic mediators in vitro. *J. Ethnopharmacol.* **129**, 203–207.
Chao, W.-W. and Lin, B.-F. (2010) Isolation and identification of bioactive compounds in Andrographis paniculata (Chuanxinlian). *Chin. Med.* **5**, 17.
Chen, H., Jones, A.D. and Howe, G.A. (2006) Constitutive activation of the jasmonate signaling pathway enhances the production of secondary metabolites in tomato. *FEBS Lett.* **580**, 2540–2546.
Chen, L.R., Chen, Y.J., Lee, C.Y. and Lin, T.Y. (2007) MeJA-induced transcriptional changes in adventitious roots of Bupleurum kaoi. *Plant Sci.* **173**, 12–24.
Chen, F., Tholl, D., Bohlmann, J. and 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.
Cherukupalli, N., Divate, M., Mittapelli, S.R., Khareedu, V.R. and Vudem, D.R. (2016) De novo assembly of leaf transcriptome in the medicinal plant Andrographis paniculata. *Front. Plant Sci.* **7**, 1203.
Criswell, J., Potter, K., Shephard, F., Beale, M.H. and Peters, R.J. (2012) A single residue change leads to a hydroxylated product from the class II diterpene cyclization catalyzed by abietadiene synthase. *Org. Lett.* **14**, 5828–5831.
Cyr, A., Wilderman, P.R., Determan, M. and Peters, R.J. (2007) A modular approach for facile biosynthesis of labdane-related diterpenes. *J. Am. Chem. Soc.* **129**, 6684–6685.
De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.
De Geyter, N., Gholami, A., Goormachtig, S. and Goossens, A. (2012) Transcriptional machineries in jasmonate-elicited plant secondary metabolism. *Trends Plant Sci.* **17**, 349–359.
Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP SignalP and related tools. *Nat. Protoc.* **2**, 953.
van der Fits, L. and Memelink, J. (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science*, **289**, 295–297.
Garg, A., Agrawal, L., Misra, R.C., Sharma, S. and Ghosh, S. (2015) Andrographis paniculata transcriptome provides molecular insights into tissue-specific accumulation of medicinal diterpenes. *BMC Genomics*, **16**, 659.
Goel, A., Kunnumakkara, A.B. and Aggarwal, B.B. (2008) Curcumin as "Curecumin": from kitchen to clinic. *Biochem. Pharmacol.* **75**, 787–809.
Haas, B.J., Papanicolaou, A., Yassour, M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512.
Hansen, N.L., Heskes, A.M., Hamberger, B., Olsen, C.E., Hallström, B.M., Andersen-Ranberg, J. and Hamberger, B. (2016) The terpene synthase gene family in Tripterygium wilfordii harbors a labdane-type diterpene synthase among the monoterpene synthase TPS-b subfamily. *Plant J.* **89**(3), 429–441.
Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. *Evolution (N. Y.)*, **22**, 160–174.
Huang, J., Pang, C., Fan, S. *et al.* (2015) Genome-wide analysis of the family 1 glycosyltransferases in cotton. *Mol. Genet. Genomics*, **290**, 1805–1818.
Inabuy, F., Fischedick, J.T., Lange, I., Hartmann, M., Srividya, N., Parrish, A.N., Xu, M., Peters, R.J. and Lange, B.M. (2017) Biosynthesis of diterpenoids in tripterygium adventitious root cultures. *Plant Physiol.* **175**(1), 92–103.
Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463.
Jian, W.W. and Wu, J.Y. (2005) Nitric oxide is involved in methyl jasmonate-induced defense responses and secondary metabolism activities of Taxus cells. *Plant Cell Physiol.* **46**, 923–930.
Jones, P. and Vogt, T. (2001) Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. *Planta*, **213**, 164–174.
Kakizaki, T., Kitashiba, H., Zou, Z., Li, F., Fukino, N. and Ohara, T. (2017) A 2-oxoglutarate-dependent dioxygenase mediates the biosynthesis of glucoraphasatin in radish. *Plant Physiol.* **173**, 1583–1593.
Kamboj, V. (2000) Herbal medicine. *Curr. Sci.* **78**, 35–39.
Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acid Res.* **28**, 27–30.
Kim, H.-J., Chen, F., Wang, X. and Rajapakse, N.C. (2006) Effect of methyl jasmonate on secondary metabolites of sweet basil (Ocimum basilicum L.). *J. Agric. Food Chem.* **54**, 2327–2332.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27**, 722–736.
Koteswara Rao, Y., Vimalamma, G., Venkata Rao, C. and Tzeng, Y.M. (2004) Flavonoids and andrographolides from Andrographis paniculata. *Phytochemistry*, **65**, 2317–2321.
Kubo, A., Arai, Y., Nagashima, S. and Yoshikawa, T. (2004) Alteration of sugar donor specificities of plant glycosyltransferases by a single point mutation. *Arch. Biochem. Biophys.* **429**, 198–203.
Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874.
Lange, B.M., Rujan, T., Martin, W. and Croteau, R. (2000) Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc. Natl Acad. Sci. USA*, **97**, 13172–13177.
Li, F. and Weng, J. (2017) Demystifying traditional herbal medicine with modern approach. *Nat. Plants*, **3**, 1–7.
Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.
Lieberman-Aiden, E., Van Berkum, N.L., Williams, L. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
Lim, J.C.W., Chan, T.K., Ng, D.S., Sagineedu, S.R., Stanslas, J. and Wong, W.F. (2012) Andrographolide and its analogues: versatile bioactive molecules for combating inflammation and cancer. *Clin. Exp. Pharmacol. Physiol.* **39**, 300–310.
Lorenc-Kukuła, K., Korobczak, A., Aksamit-Stachurska, A., Kostyń, K., Lukaszewicz, M. and Szopa, J. (2004) Glucosyltransferase: the gene arrangement and enzyme function. *Cell. Mol. Biol. Lett.* **9**, 935–946.
Lu, X., Zhang, L., Zhang, F., Jiang, W., Shen, Q., Zhang, L., Lv, Z., Wang, G. and Tang, K. (2013) AaORA, a trichome-specific AP2/ERF transcription factor of Artemisia annua, is a positive regulator in the artemisinin biosynthetic pathway and in disease resistance to Botrytis cinerea. *New Phytol.* **198**, 1191–1202.
Ma, X.C., Gou, Z.P., Wang, C.Y., Yao, J.H., Xin, X.L., Lin, Y. and Liu, K.X. (2010) A new ent-labdane diterpenoid lactone from Andrographis paniculata. *Chin. Chem. Lett.* **21**, 587–589.
Ma, Y., Yuan, L., Wu, B., Li, X., Chen, S. and Lu, S. (2012) Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in Salvia miltiorrhiza. *J. Exp. Bot.* **63**, 2809–2823.
Mackenzie, P.I., Bock, K.W., Burchell, B., Guillemette, C., Ikushiro, S., Iyanagi, T., Miners, J.O., Owens, I.S. and Nebert, D.W. (2005) Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet. Genomics*, **15**, 677–685.
Mafu, S., Potter, K.C., Hillwig, M.L., Schulte, S., Criswell, J.D. and Peters, R.J. (2015) Efficient heterocyclisation by (di)terpene synthases. *Chem. Commun.* **51**, 13485–13487.

© 2018 The Authors

Misra, R.C., Garg, A., Roy, S., Chanotiya, C.S., Vasudev, P.G. and Ghosh, S. (2015) Involvement of an ent-copalyl diphosphate synthase in tissue-specific accumulation of specialized diterpenes in Andrographis paniculata. *Plant Sci.* **240**, 50–64.

Nelson, D. and Werck-Reichhart, D. (2011) A P450-centric view of plant evolution. *Plant J.* **66**, 194–211.

Nuetzmann, H.-W. and Osbourn, A. (2015) Regulation of metabolic gene clusters in Arabidopsis thaliana. *New Phytol.* **205**, 503–510.

Pal, S. and Shukla, Y. (2003) Herbal medicine: current status and the future. *Asian Pac. J. Cancer Prev.* **4**, 281–288.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA*, **101**, 9903–9908.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551.

Peters, R.J. (2010) Two rings in them all: the labdane-related diterpenoids. *Nat. Prod. Rep.* **27**, 1521–1530.

Peters, R.J., Flory, J.E., Jetter, R., Ravn, M.M., Lee, H.J., Coates, R.M. and Croteau, R.B. (2000) Abietadiene synthase from grand fir (Abies grandis): characterization and mechanism of action of the "pseudomature" recombinant enzyme. *Biochemistry*, **39**, 15592–15602.

Peters, R.J., Ravn, M.M., Coates, R.M. and Croteau, R.B. (2001) Bifunctional abietadiene synthase: free diffusive transfer of the (+)-copalyl diphosphate intermediate between two distinct active sites. *J. Am. Chem. Soc.* **123**, 8974–8978.

Pholphana, N., Rangkadilok, N., Saehun, J., Ritruechai, S. and Satayavivad, J. (2013) Changes in the contents of four active diterpenoids at different growth stages in Andrographis paniculata (Burm.f.) Nees (Chuanxinlian). *Chin. Med.* **8**, 2.

Poppenberger, B., Fujioka, S., Soeno, K. *et al.* (2005) The UGT73C5 of Arabidopsis thaliana glucosylates brassinosteroids. *Proc. Natl Acad. Sci. USA*, **102**, 15253–15258.

Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Potter, K., Criswell, J., Zi, J., Stubbs, A. and Peters, R.J. (2014) Novel product chemistry from mechanistic analysis of ent-copalyl diphosphate synthases from plant hormone biosynthesis. *Angew. Chem. Int. Ed. Engl.* **53**, 7198–7202.

Prisic, S., Xu, J., Coates, R.M. and Peters, R.J. (2007) Probing the role of the DXDD motif in class II diterpene cyclases. *Chembiochem*, **8**, 869–874.

Raskin, I., Ribnicky, D.M., Komarnytsky, S. *et al.* (2002) Plants and human health in the twenty-first century. *Trends Biotechnol.* **20**, 522–531.

Reyes, B.A.S., Bautista, N.D., Tanquilut, N.C. *et al.* (2006) Anti-diabetic potentials of Momordica charantia and Andrographis paniculata and their effects on estrous cyclicity of alloxan-induced diabetic rats. *J. Ethnopharmacol.* **105**, 196–200.

Richman, A., Swanson, A., Humphrey, T., Chapman, R., McGarvey, B., Pocs, R. and Brandle, J. (2005) Functional genomics uncovers three glucosyltransferases involved in the synthesis of the major sweet glucosides of Stevia rebaudiana. *Plant J.* **41**, 56–67.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Ross, J., Li, Y., Lim, E. and Bowles, D.J. (2001) Higher plant glycosyltransferases. *Genome Biol.* **2**, REVIEWS3004.

Roy Upton, R.H. (2015) Traditional herbal medicine, pharmacognosy, and pharmacopoeial standards: A discussion at the crossroads. In *Evidence-Based Validation of Herbal Medicine* (Mukherjee, R., ed.) ISBN 987-0-12-800874-4. http://dx.doi.org/10.1016/B978-0-12-800874-4.00003-9.

Rushton, P.J., Somssich, I.E., Ringler, P. and Shen, Q.J. (2010) WRKY transcription factors. *Trends Plant Sci.* **15**, 247–258.

Sato, S., Tabata, S., Hirakawa, H. *et al.* (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Sayama, T., Ono, E., Takagi, K. *et al.* (2012) The Sg-1 glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. *Plant Cell*, **24**, 2123–2138.

Sc, M. (2014) Establishment of hairy root for andrographolide production in Andrographis paniculata.

Schluttenhofer, C. and Yuan, L. (2015) Regulation of specialized metabolism by WRKY transcription factors. *Plant Physiol.* **167**, 295–306.

Schweizer, F., Fernández-Calvo, P., Zander, M. *et al.* (2013) Arabidopsis basic helix-loop-helix transcription factors MYC2, MYC3, and MYC4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. *Plant Cell*, **25**, 3117–3132.

Sheeja, K., Shihab, P.K. and Kuttan, G. (2006) Antioxidant and anti-inflammatory activities of the plant Andrographis paniculata Nees. *Immunopharmacol. Immunotoxicol.* **28**, 129–140.

Shen, Q., Lu, X., Yan, T. *et al.* (2016) The jasmonate-responsive AaMYC2 transcription factor positively regulates artemisinin biosynthesis in Artemisia annua. *New Phytol.* **210**, 1269–1281.

Singha, P.K., Roy, S. and Dey, S. (2003) Antimicrobial activity of Andrographis paniculata. *Fitoterapia*, **74**, 692–694.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Subramanian, R., Asmawi, M.Z. and Sadikun, A. (2012) A bitter plant with a sweet future? A comprehensive review of an oriental medicinal plant: andrographis paniculata. *Phytochem. Rev.* **11**, 39–75.

Sudhakaran, M. (2012) Botanical pharmacognosy of Andrographis paniculata (Burm. F.) Wall. *Ex. Nees. Pharmacogn. J.* **4**, 1–10.

Swaminathan, S., Morrone, D., Wang, Q., Fulton, D.B. and Peters, R.J. (2009) CYP76M7 is an ent-Cassadiene C11-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell*, **21**, 3315–3325.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M. and Paterson, A.H. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954.

Tarailo-Graovac, M. and Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, **5**(1), 4.10.1–4.10.14.

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Van der Fits, L. and Memelink, J. (2001) The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. *Plant J.* **25**, 43–53.

Van Moerkercke, A., Steensma, P., Schweizer, F. *et al.* (2015) The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in Catharanthus roseus. *Proc. Natl Acad. Sci. USA*, **112**, 201504951.

Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van De Peer, Y. and Geuten, K. (2012) Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-BOX gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806.

Vogt, T. and Jones, P. (2000) Glycosyltransferases in plant natural product synthesis: characterization of a supergene family. *Trends Plant Sci.* **5**, 380–386.

Vranová, E., Coman, D. and Gruissem, W. (2013) Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annu. Rev. Plant Biol.* **64**, 665–700.

Wang, G.C., Wang, Y., Williams, I.D., Sung, H.H.Y., Zhang, X.Q., Zhang, D.M., Jiang, R.W., Yao, X.S. and Ye, W.C. (2009) Andrographolactone, a unique diterpene from Andrographis paniculata. *Tetrahedron Lett.* **50**, 4824–4826.

Wang, H., Ma, C., Li, Z., Ma, L., Wang, H., Ye, H., Xu, G. and Liu, B. (2010) Effects of exogenous methyl jasmonate on artemisinin biosynthesis and secondary metabolites in Artemisia annua L. *Ind. Crops Prod.* **31**, 214–218.

Wang, Q., Hillwig, M.L., Wu, Y. and Peters, R.J. (2012a) CYP701A8: a rice ent-kaurene oxidase paralog diverted to more specialized diterpenoid metabolism. *Plant Physiol.* **158**, 1418–1425.

Wang, Y., Tang, H., Debarry, J.D. *et al.* (2012b) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wang, L., Yu, S., Tong, C. *et al.* (2014) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* **15**, R39.

Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**(3), 543–548.

© 2018 The Authors

World Health Organization (2002) Flos caryophylli. *WHO Monogr.* **2**, 45–52.
Xu, Z. and Song, J. (2017) The 2-oxoglutarate-dependent dioxygenase superfamily participates in tanshinone production in Salvia miltiorrhiza. *J. Exp. Bot.* **68**, 2299–2308.
Xu, H., Song, J., Luo, H. *et al.* (2016) Analysis of the genome sequence of the medicinal plant Salvia miltiorrhiza. *Mol. Plant.* **9**, 949–952.
Yamada, Y., Koyama, T. and Sato, F. (2011) Basic helix-loop-helix transcription factors and regulation of alkaloid biosynthesis. *Plant Signal. Behav.* **6**, 1627–1630.
Yamaguchi, S., Sun, T.P., Kawaide, H. and Kamiya, Y. (1998) The GA2 locus of Arabidopsis thaliana encodes ent-kaurene synthase of gibberellin biosynthesis. *Plant Physiol.* **116**, 1271–1278.
Yamamura, C., Mizutani, E., Okada, K. *et al.* (2015) Diterpenoid phytoalexin factor, a bHLH transcription factor, plays a central role in the biosynthesis of diterpenoid phytoalexins in rice. *Plant J.* **84**, 1100–1113.
Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
Yin, Q., Shen, G., Chang, Z., Tang, Y., Gao, H. and Pang, Y. (2017) Involvement of three putative glucosyltransferases from the UGT72 family in flavonol glucoside/rhamnoside biosynthesis in Lotus japonicus seeds. *J. Exp. Bot.* **68**, 597–612.
Yonekura-Sakakibara, K. and Hanada, K. (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J.* **66**, 182–193.
Yu, Z.X., Li, J.X., Yang, C.Q., Hu, W.L., Wang, L.J. and Chen, X.Y. (2012) The jasmonate-responsive AP2/ERF transcription factors AaERF1 and AaERF2 positively regulate artemisinin biosynthesis in Artemisia annua L. *Mol. Plant*, **5**, 353–365.
Zhang, H., Hedhili, S., Montiel, G., Zhang, Y., Chatel, G., Pré, M., Gantet, P. and Memelink, J. (2011) The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *Plant J.* **67**, 61–71.
Zhou, Y., Sun, W. and Chen, J. (2016) SmMYC2a and SmMYC2b played similar but irreplaceable roles in regulating the biosynthesis of tanshinones and phenolic acids in Salvia miltiorrhiza. *Sci. Rep.* **6**, 22852.
Zi, J., Mafu, S. and Peters, R.J. (2014a) To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annu. Rev. Plant Biol.* **65**, 259–286.
Zi, J., Matsuba, Y., Hong, Y.J., Jackson, A.J., Tantillo, D.J., Pichersky, E. and Peters, R.J. (2014b) Biosynthesis of lycosantalonol, a cis-prenyl derived diterpenoid. *J. Am. Chem. Soc.* **136**, 16951–16953.

© 2018 The Authors
The Plant Journal © 2018 John Wiley & Sons Ltd,