Plant Communications
Research Article

CellPress
Partner Journal

The origin and evolution of the diosgenin biosynthetic pathway in yam

Jian Cheng[1,6], Jing Chen[1,2,6], Xiaonan Liu[1,6], Xiangchen Li[3,6], Weixiong Zhang[4,6], Zhubo Dai[1], Lina Lu[1], Xiang Zhou[5], Jing Cai[4,*], Xueli Zhang[1,*], Huifeng Jiang[1,*] and Yanhe Ma[1]

[1]Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]College of Life Science, Northwest A&F University, Yangling, Shaanxi 712100, China

[4]Research Center for Ecology and Environmental Sciences, Northwestern Polytechnical University, Xian, China

[5]Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi 330004, China

[6]These authors contributed equally to this article.

***Correspondence: Jing Cai (jingcai@nwpu.edu.cn), Xueli Zhang (zhang_xl@tib.cas.cn), Huifeng Jiang (jiang_hf@tib.cas.cn)**

https://doi.org/10.1016/j.xplc.2020.100079

ABSTRACT

Diosgenin, mainly produced by *Dioscorea* species, is a traditional precursor of most hormonal drugs in the pharmaceutical industry. The mechanisms that underlie the origin and evolution of diosgenin biosynthesis in plants remain unclear. After sequencing the whole genome of *Dioscorea zingiberensis*, we revealed the evolutionary trajectory of the diosgenin biosynthetic pathway in *Dioscorea* and demonstrated the *de novo* biosynthesis of diosgenin in a yeast cell factory. First, we found that P450 gene duplication and neo-functionalization, driven by positive selection, played important roles in the origin of the diosgenin biosynthetic pathway. Subsequently, we found that the enrichment of diosgenin in the yam lineage was regulated by CpG islands, which evolved to regulate gene expression in the diosgenin pathway and balance the carbon flux between the biosynthesis of diosgenin and starch. Finally, by integrating genes from plants, animals, and yeast, we heterologously synthesized diosgenin to 10 mg/l in genetically-engineered yeast. Our study not only reveals the origin and evolutionary mechanisms of the diosgenin biosynthetic pathway in *Dioscorea*, but also introduces an alternative approach for the production of diosgenin through synthetic biology.

Keywords: metabolic engineering, genomic evolution, diosgenin, synthetic biology

Cheng J., Chen J., Liu X., Li X., Zhang W., Dai Z., Lu L., Zhou X., Cai J., Zhang X., Jiang H., and Ma Y. (2021). The origin and evolution of the diosgenin biosynthetic pathway in yam. Plant Comm. **2**, 100079.

INTRODUCTION

The plant kingdom collectively produces hundreds of thousands of specialized metabolites (Dixon, 2001; Osbourn and Lanzotti, 2009) that have been used as traditional medicines for preventing and treating different kinds of diseases (Balandrin et al., 1985; Dudareva et al., 2006). Even today, natural products from plants play important roles in modern drugs, and more than 70% of modern drugs have been directly or indirectly derived from specialized plant metabolites over the past two decades (De Luca et al., 2012). Unfortunately, specialized metabolites are usually present at very low concentrations in plants; for example, there is approximately 0.05% taxol in the bark of yew (Nadeem et al., 2002), and only 0.01% lycopene in fresh tomatoes (Alda et al., 2009). Due to the extremely wide diversity of metabolites and the highly complex network of metabolic pathways in plants, specialized metabolites are usually produced only in particular tissues at specific times (Maeda, 2019; Wang et al., 2019). Therefore, the accumulation of specialized metabolites in plants often requires particular environmental conditions over a long period of time. The use of specialized plant metabolites for industrial purposes is still largely limited by low yields (Wang et al., 2019).

Due to their enrichment in specialized metabolites, more than 10 000 medicinal plants have been used in Traditional Chinese Medicine for thousands of years (Tang and Eisenbrand, 1993). For example, *Dioscorea zingiberensis*, an important medicinal herb in Traditional Chinese Medicine since the Han dynasty, can accumulate 2%–16% diosgenin in the rhizomes (Huang et al., 2008). Diosgenin is an important precursor for the production of many steroidal drugs, including antioxidants, anti-inflammatories, sex hormones, steroids, cortisone,

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

Genome parameters	Values
Total size	480.27 Mb
Scaffold N50	44.55 Mb
Scaffold L50	5 scaffolds
Scaffold N90	34.33 Mb
Scaffold L90	10 scaffolds
Number of scaffolds	16 452
Genome GC content	35.20%
Total size of repetitive sequences	182.5 M
Long terminal repeat retrotransposons	137.5 M
DNA transposons	19.2 M
Long interspersed nuclear elements	21.9 M
Number of coding genes	26 022
Average gene length	1139 bp
Number of non-coding RNAs	2232

Table 1. Statistics of genome assembly, repetitive sequences, and gene annotation.

contraceptives, fertility control compounds, and anabolic agents (Fernandes et al., 2003; Wang et al., 2007; He et al., 2012). In *Dioscorea*, the biosynthesis of diosgenin from the precursor cholesterol (Sonawane et al., 2016) is catalyzed by two P450 enzymes: C-16,22-dihydroxylase and C-26 hydroxylase (Christ et al., 2019). *Dioscorea* appears to have evolved genes that overcome the bottleneck of complex metabolic network regulation to improve diosgenin production. However, the origin and evolutionary mechanisms of the diosgenin biosynthetic pathway in *Dioscorea* remain poorly understood.

Here, by scrutinizing the genome of *D. zingiberensis*, we first reveal the origin and evolution of the diosgenin biosynthetic pathway in yam at the genomic level. Then, we biosynthesize diosgenin *de novo* in engineered yeast by integrating genes from plants, animals, and yeast.

RESULTS

Positive selection has driven the origin of the diosgenin biosynthetic pathway

To characterize the diosgenin biosynthetic pathway in *D. zingiberensis*, we first obtained a 480.3 Mb high-quality genome sequence at the chromosomal level (Table 1) by combining reads from the Illumina HiSeq (205 Gb, 393×), PacBio SMRT (21 Gb, 40×), and Hi-C (62 Gb clean data) platforms (Supplemental Table 1). We fully annotated 26 022 protein-coding genes (Supplemental Tables 2, 3, 4, and 5), which covered 97% of core genes based on CEGMA (Parra et al., 2007) and BUSCO (Simão et al., 2015) assessments (Supplemental Table 6). A recent study reported that diosgenin in *Paris polyphylla* and *Trigonella foenum–graecum* is independently synthesized from cholesterol by two P450 genes (*PpCYP90G4* in combination with *PpCYP94D108*, *PpCYP94D109*, or *PpCYP72A616*, or *TfCYP90B50* in combination with *TfCYP72A613* or *TfCYP82J17*), corresponding to steroid C-16,22-dihydroxylase and C-26 hydroxylase (Christ et al., 2019). These genes were used as blastp queries against the proteins in *D. zingiberensis*, and only orthologs of *PpCYP90G4* and *PpCYP94D108* were identified in *D. zingiberensis* (*DzinCYP90G6* and *DzinCYP94D144*) (Figure 1A, Supplemental Table 7).

To clarify the origin of the diosgenin biosynthetic pathway, we analyzed genomic and transcriptomic data from all sequenced species in the *Petrosaviidae* clade. Based on sequence similarity and intron distribution (Supplemental Figure 1), we found that *DzinCYP90G6* and its duplicated copy *DzinCYP90B71* originated from the common ancestor *DzinCYP90B77* (Supplemental Figure 2), which is conserved in almost all higher plants and functions as a campesterol C-22 hydroxylase in the brassinosteroid biosynthetic pathway (Satomi et al., 2010). Interestingly, we found that *DzinCYP90B71*, *DzinCYP94D144*, and *DzinCYP90G6* formed a gene cluster (Figure 1B) that was retained only in the four studied species of the *Petrosaviidae* clade (*D. zingiberensis*, *D. rotundata*, *Musa acuminata*, and *Phoenix dactylifera*). Moreover, we only detected diosgenin in species that contained this gene cluster (Figure 1C). Although some species contained some of the genes in the cluster, we did not detect diosgenin in the other seven monocots. These results demonstrated that the P450 gene cluster containing *DzinCYP90B71*, *DzinCYP94D144*, and *DzinCYP90G6* played a key role in the origination of the diosgenin biosynthesis pathway.

The function of diosgenin in plant defense against pathogens, pests, and herbivores implies that diosgenin biosynthesis is important for species adaptation (Moses et al., 2014). We calculated the evolutionary rates of the CYP90G subfamily and observed significant positive selection in the CYP90G branch based on lineage-specific dN/dS ratios (Figure 1D, Supplemental Table 8). This indicated that gene duplication and subsequent neo-functionalization of CYP90G was critical for the origin of the diosgenin biosynthetic pathway in the *Petrosaviidae* clade. Another P450 gene branch in the cluster, CYP94X, which is also a *Petrosaviidae*-specific branch (Supplemental Figure 3), may have arisen from CYP94Y by

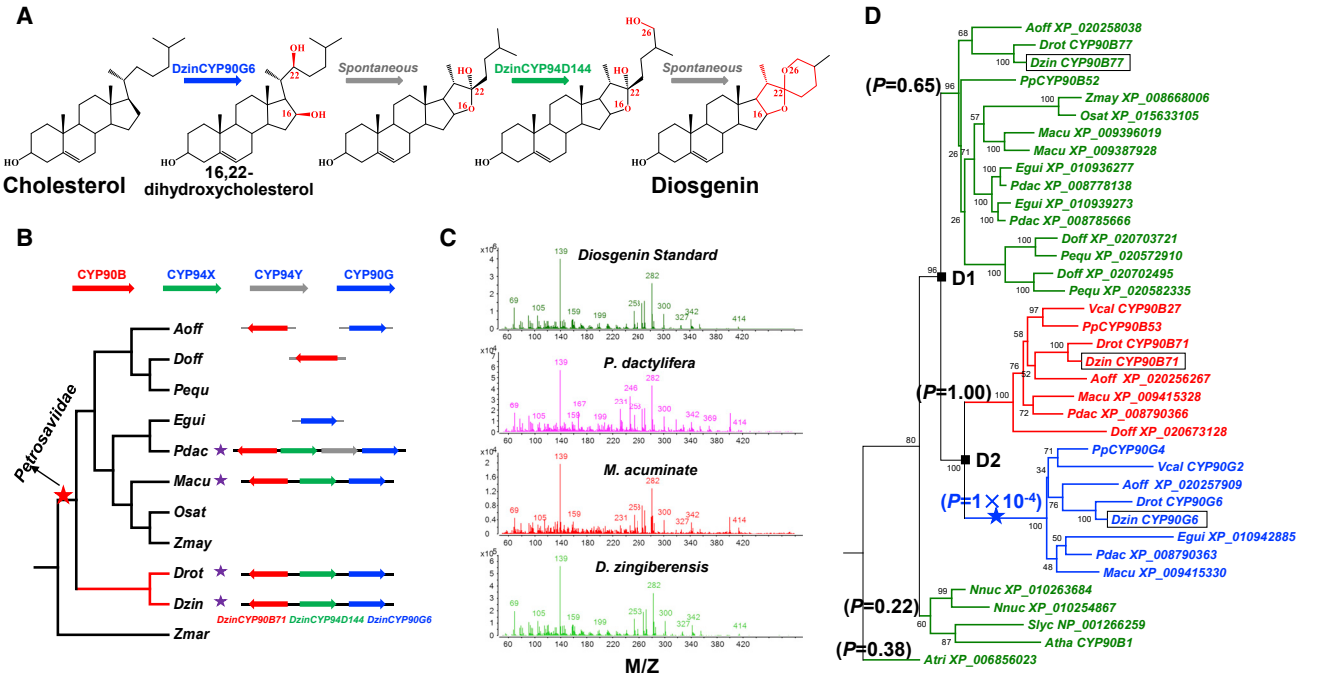

Figure 1. Origin and evolution of the diosgenin biosynthetic pathway.

(A) Two P450 genes for diosgenin biosynthesis, *DzinCYP90G6* and *DzinCYP94D144*, which are orthologs of *PpCYP90G4* and *PpCYP94D108* that catalyze the production of diosgenin from cholesterol, were identified in *D. zingiberensis*.

(B) The distribution of the P450 cluster in 11 monocots. The species names are abbreviated as follows: *A. officinalis* (Aoff), *D. officinale* (Doff), *P. equestris* (Pequ), *E. guineensis* (Egui), *P. dactylifera* (Pdac), *M. acuminata* (Macu), *O. sativa* (Osat), *Z. mays* (Zmay), *D. rotundata* (Drot), *D. zingiberensis* (Dzin), and *Z. marina* (Zmar). Both CYP94X and CYP94Y branches belong to the CYP94D subfamily.

(C) GC-MS analysis of diosgenin in *D. zingiberensis*, *M. acuminata*, and *P. dactylifera*.

(D) The phylogenetic relationships between the CYP90B and CYP90G subfamilies. Except for genes with known CYP nomenclatures, we have used NCBI accession IDs. D1 and D2 represent two possible duplication events in the evolution of three genes. The *P*-values represent the test significance for the detection of positive selection corresponding to five branches of the CYP90 family.

tandem duplication because both subfamilies are present in the *P. dactylifera* cluster, and CYP94Y has been lost in other species (Figure 1B). We speculate that the P450 gene cluster originated after the divergence of *Petrosaviidae* from *Alismatales* approximately 120–138 MYA, initiating biosynthesis of diosgenin in species of the *Petrosaviidae* clade.

Evolution of the diosgenin biosynthetic pathway in yam

Unlike *D. zingiberensis*, *D. rotundata* contains only a trace amount of diosgenin and is the main source of starch in tropical Africa (Mignouna et al., 2009; Tamiru et al., 2017). To characterize differences in metabolic flux between the two species, we analyzed the expression of genes involved in diosgenin and starch biosynthesis. Most genes (11 of 12) in the diosgenin biosynthetic pathway had higher expression in rhizomes than in leaves of *D. zingiberensis*, but this pattern was not present in *D. rotundata* (Supplemental Figure 4). The expression of seven diosgenin biosynthetic genes in *D. zingiberensis* was at least twice as high as in *D. rotundata* (Figure 2A, left panel). By contrast, 16 genes related to starch synthesis, storage, and degradation (Zhixi et al., 2009) in the rhizome had higher expression levels in *D. rotundata* than in *D. zingiberensis* (Figure 2A, right panel). Thirteen of these 16 genes had expression levels at least twice as high in *D. rotundata* as in *D. zingiberensis* (Figure 2A, right panel). Four genes (i.e., *AGPL*, ADP-glucose pyrophosphorylase large subunit; *AGPS*, ADP-glucose pyrophosphorylase small subunit; *PHOH*, starch cytosolic phosphorylase; and *DPE2*, disproportionating enzyme) exhibited 5- to 10-fold differences in expression, and two genes (i.e., *SBE1*, starch branching enzyme; and *PHOL*, starch plastidial phosphorylase) exhibited expression differences greater than 20-fold. The expression of these genes in *D. rotundata* was clearly higher than that in the other nine monocots (Supplemental Figure 5). These results demonstrated that the evolution of gene expression balanced the biosynthesis of diosgenin with that of starch in different yams.

CpG density at promoter regions often modifies the methylation level and thereby contributes to transcriptional regulation of downstream genes (Deaton and Bird, 2011). To further explore the mechanisms underlying different patterns of gene expression, we analyzed CpG distribution at the promoter region of genes in both pathways. Interestingly, we found that the CpG density upstream of diosgenin biosynthesis genes was higher in *D. rotundata* than in *D. zingiberensis* (Figure 2B). The opposite CpG pattern was observed for orthologous genes in the starch biosynthesis pathway. Differences in CpG density upstream of diosgenin and starch genes in the two yam species probably contributed to differences in gene expression, in turn leading to drastic differences in the diosgenin and starch content of the two species.

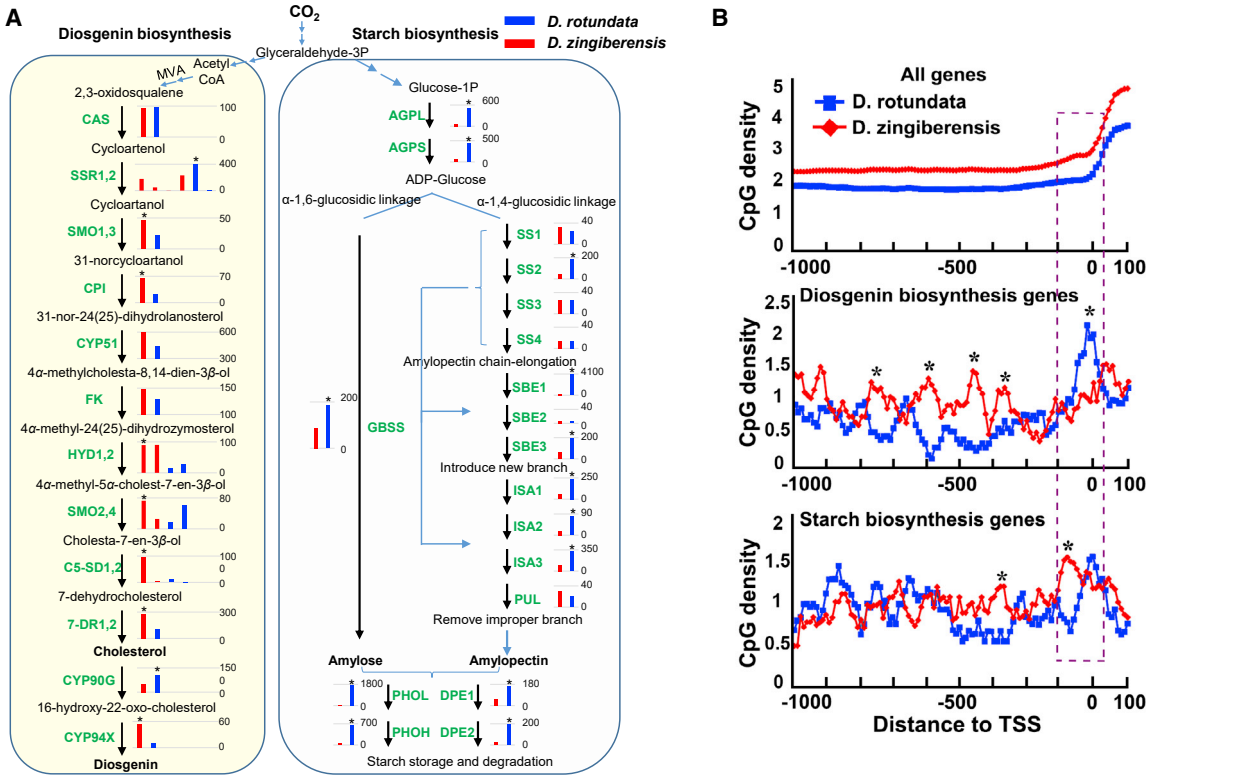

Figure 2. Evolution of the diosgenin biosynthesis pathway and comparison with the starch biosynthesis pathway in *D. zingiberensis* and *D. rotundata*.

(A) Gene expression analysis for the diosgenin and starch biosynthesis pathways. The left frame shows the diosgenin biosynthesis pathway, and the right frame shows genes involved in starch synthesis, storage, and degradation. The bar charts show the corresponding gene expression levels (FPKM) in rhizome tissue. The red and blue colors represent gene expression in *D. zingiberensis* and *D. rotundata*, respectively. More than one bar with the same color in a bar chart indicates that the gene has multiple copies in a species. An asterisk indicates that a gene's expression in one species is at least twice as high as that of its ortholog in the other species. CAS, cycloartenol synthase; SSR, sterol side chain reductase; SMO, C-4 sterol methyl oxidase; CPI, cyclopropylsterol isomerase; CYP51, sterol C-14 demethylase; FK, sterol C-14 reductase; HYD, sterol 8,7 isomerase; C5-SD, sterol C-5(6) desaturase 2; 7-DR, 7-dehydrocholesterol reductase 2. AGPL, ADP-glucose pyrophosphorylase large subunit; AGPS, ADP-glucose pyrophosphorylase small subunit; GBSS, granule-bound starch synthase; SS, soluble starch synthase; SBE, starch branching enzyme; ISA, isoamylase; PUL, pullulanase; DBE, starch debranching enzyme; PHO, starch plastidial phosphorylase; DPE, Disproportionating enzyme.

(B) Comparison of CpG density in the upstream regions of diosgenin and starch biosynthesis genes. The three charts show the average CpG density at the promoter regions of all genes, diosgenin biosynthesis genes, and starch biosynthesis genes, respectively. The diosgenin and starch biosynthesis gene lists are the same as in **(A)**. Asterisks indicate locations at which CpG density differs significantly between the two species.

Biosynthesis of diosgenin in yeast by metabolic engineering

Because the biosynthetic pathway of cholesterol in animal cells is more efficient than that in plants, we first combined genes from animals to construct a cholesterol-producing yeast chassis strain, as in the original work by Souza et al. (2011) (Figure 3A and 3B). The previously engineered *Saccharomyces cerevisiae* strain BY-T3 was selected as the starting strain because it could produce large amounts of 2,3-oxidosqualene, a common precursor for triterpenoid biosynthesis (Dai et al., 2019). First, the native alcohol acetyltransferase (*ATF2*) gene was deleted in the starting strain BY-T3 to avoid acetylation of the steroid C3-hydroxyl groups (Florence Ménard et al., 2003) (Supplemental Figure 6). Then, the dehydrocholesterol 7-reductase (*DHCR7*) and dehydrocholesterol 24-reductase (*DHCR24*) genes from *Danio rerio* were synthesized with codon optimization and transformed together into the δATF2 strain (Souza et al., 2011) (Figure 3A and 3B). In addition, the native ergosterol synthesis pathway in yeast was blocked by deleting the *ERG6* gene to decrease sterol competition and improve cholesterol production (Figure 3A and 3B), and the galactose-inducible promoter *Gal1* was used to control the expression of *DrDHCR24* and alleviate cell damage by cholesterol synthesis. Finally, a cholesterol-producing yeast chassis strain DG-Cho was constructed and shown to produce 16 mg/l cholesterol after 120 h of shake flask fermentation (Supplemental Figures 7 and 8).

Next, *DzinCYP90G6*, *DzinCYP94D144*, and *VvCPR* were co-expressed in the cholesterol-producing yeast strain DG-Cho to obtain strain DG002, which was able to convert cholesterol to diosgenin to approximately 3.2 mg/l (Figure 3C, Supplemental Figure 9). In addition, steroid C-16,22-dihydroxylase and C-26 hydroxylase genes from different species, including *D. zingiberensis*, *P. polyphylla, T. foenum–graecum*, and *Veratrum californicum*, were investigated for their combined effects on diosgenin production. After comparison of various steroid C-16,22-dihydroxylases and C-26 hydroxylases, *DzinCYP90G6* was identified as the best steroid C-16,22-dihydroxylase, and

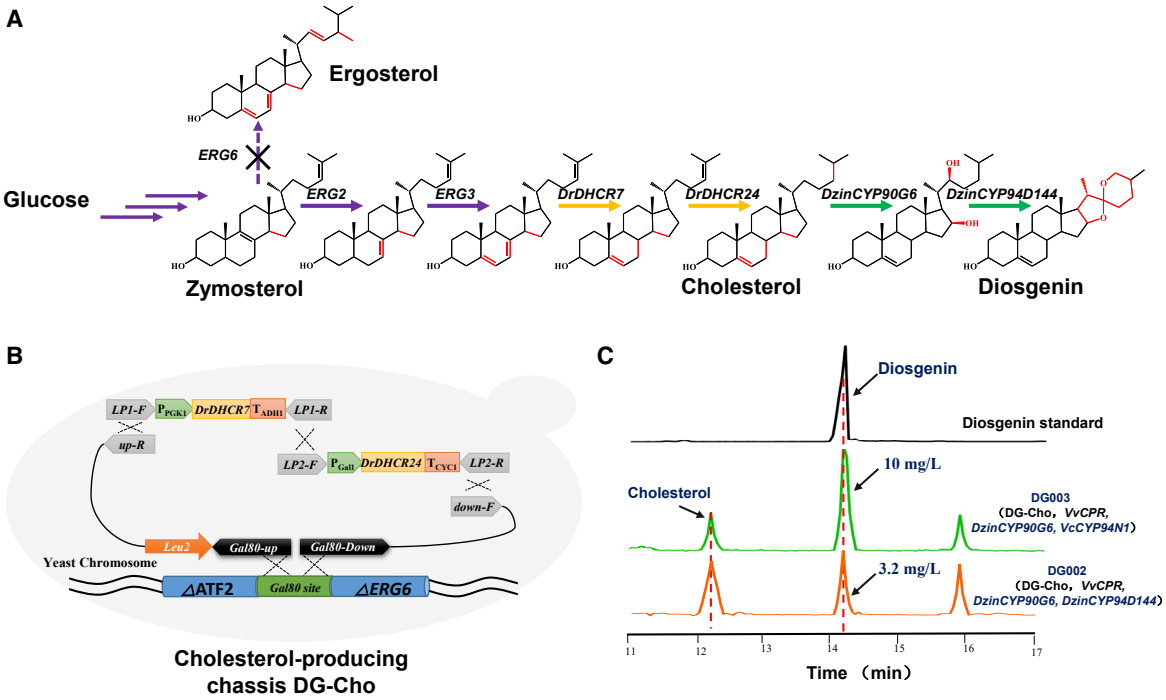

Figure 3. *De Novo* biosynthesis of diosgenin in a yeast platform.

(A) Reconstruction of the diosgenin biosynthesis pathway in yeast. The purple steps represent endogenous genes from yeast, and the orange and green steps represent the heterologous genes from *D. rerio* and *D. zingiberensis*, respectively.

(B) Construction of the cholesterol-producing platform in yeast.

(C) GC-MS profiles for diosgenin biosynthesis in engineered yeast. Each colored chromatograph corresponds to the following: black, diosgenin standard; green, DG003 for diosgenin biosynthesis using *DzinCYP90G6* and *VcCYP94N1*; red, DG002 for diosgenin biosynthesis using *DzinCYP90G6* and *DzinCYP94D144*.

VcCYP94N1 from *V. californicum* was identified as the best steroid C-26 hydroxylase. The combined expression of *DzinCYP90G6* and *VcCYP94N1* in the cholesterol-producing strain DG-Cho (DG003) produced a diosgenin titer of 10 mg/l after a 120-h fermentation (Figure 3C), which was the highest diosgenin titer to date.

DISCUSSION

The plant kingdom collectively produces hundreds of thousands of specialized metabolites. During the evolution of plants, novel compounds continuously arise in specific lineages, potentially as the outcome of co-evolution with natural enemies (Pichersky and Raguso, 2016). Therefore, it is possible to investigate the evolutionary trajectory of chemodiversity in plants by comparative genomics of close relatives (Weng and Noel, 2012). Here, we used comparative genomic analysis to uncover three key evolutionary events in the origin and evolution of the diosgenin biosynthetic pathway in *D. zingiberensis* (Supplemental Figure 10). First, we speculated that the emergence of a novel compound (diosgenin) would undergo strong positive selection, resulting in the P450 gene cluster for diosgenin biosynthesis. Subsequently, we observed that some genes involved in cholesterol biosynthesis had duplicated and were highly expressed, like those in *Solanum lycopersicum* (Supplemental Figure 11), suggesting that yam may have independently evolved a cholesterol biosynthetic pathway by *Dioscorea*-specific duplications for the optimization of precursor biosynthesis. A similar phenomenon has also been observed in the opium poppy, which assembled the benzylisoquinoline alkaloid gene cluster containing the *STORR* gene fusion essential for morphinan biosynthesis prior to a whole-genome duplication (WGD) event (Guo et al., 2018). Finally, the diosgenin biosynthetic pathway was improved by weakening the competing starch biosynthesis pathway, which may have been dramatically strengthened by regulation through epigenetic modification. Our results not only decode the evolutionary trajectory of the diosgenin biosynthetic pathway but also provide insights into the mechanisms of origin of chemodiversity in plants (Mignouna et al., 2009).

Dioscorea is a large genus of more than 600 species worldwide, some of which are important food or drug sources (Govaerts and Saunders, 2007). Because *Dioscorea* exhibits features of both dicots and monocots, it has the potential to fill gaps in our knowledge of the plant biology and evolution of these two taxa (Mignouna et al., 2009). The high-quality *D. zingiberensis* genome sequence presented here can also serve as a reference genome for other species in the genus *Dioscorea*. For example, the *Dioscorea*-specific WGD event has been neglected in previous studies due to the lack of a reference genome (Ren et al., 2018). According to our genome analysis, approximately 35.9% syntenic regions were observed in the *D. zingiberensis* genome compared with 15.8% in *D. rotundata* (Supplemental Table 9, Supplementary Methods). The distribution of synonymous substitution rates (K_S) in 2018

paralogous genes from syntenic regions exhibited a peak at K_S = 0.82 in *D. zingiberensis* (Supplemental Figure 12), indicating that the *Dioscorea*-specific WGD event may have occurred after divergence from the *Petrosaviidae* but before the separation of *D. zingiberensis* and *D. rotundata* (K_S = 0.4). The date of the WGD was estimated at approximately 64 MYA (Supplemental Figure 13, Supplementary Methods), close to the Cretaceous-Paleogene extinction event (Renne et al., 2013). The WGD event provided plentiful genetic materials for *Dioscorea* species to survive across the Cretaceous-Paleogene boundary (Kevin et al., 2014) and also contributed to the evolution of specialized metabolites (Guo et al., 2014; Jing et al., 2015; Unver et al., 2017; Li et al., 2019).

The worldwide pharmaceutical industry market for steroidal drugs has reached US$ 4–8 billion (Bai et al., 2015; Yosef Al Jasem et al., 2014). However, the conventional methods for isolation and purification of diosgenin from *D. zingiberensis* have many disadvantages, such as low efficiency and high contamination (Yang et al., 2015; Mafalda et al., 2016). Therefore, it is valuable to characterize the diosgenin biosynthesis pathway and to supply diosgenin in a cost-effective and environmentally friendly manner (Liu et al., 2018). To improve the efficiency of diosgenin biosynthesis, we constructed a chimeric pathway by integrating genes from zebrafish, yeast, *Veratrum*, and *Dioscorea*. The engineered yeast with two genes from zebrafish accumulated cholesterol at 16 mg/l, which is significantly higher than that reported in studies using the plant cholesterol pathway (Sonawane et al., 2016; Christ et al., 2019), indicating that genes from the animal cholesterol pathway are more efficient than those from plants. Furthermore, the use of two P450 genes from *Veratrum* and *Dioscorea* to convert cholesterol to diosgenin improved the diosgenin yield three-fold and the conversion rate from cholesterol to diosgenin by 60%. It would be feasible to further enhance the production of diosgenin by optimizing the key enzymes that convert cholesterol to diosgenin. In summary, our results not only contribute to decoding the evolutionary trajectory of the diosgenin biosynthetic pathway but also provide insights into enhancing the production of diosgenin by metabolic engineering in microbes.

METHODS

Constructing the phylogenetic tree of the CYP90 family

Multiple alignment of protein sequences, including the CYP90B and CYP90G subfamilies, was performed using MAFFT v7.394 (Katoh and Standley, 2013) with default parameters and back-translated using PAL2NAL v14 (Suyama et al., 2006). The ML phylogenetic trees were constructed from the cDNA alignment with the software RAxML v8.2.4 (Stamatakis, 2014) using the GTR + Γ + I substitution model and mapping the percentage of 500 rapid bootstraps to the best-scoring ML tree with *Amborella trichopoda* as an outgroup.

Molecular evolution analyses of the CYP90 family

To evaluate the variation in selection pressures over different P450 clades, the free ratio model of the codeml algorithm (Yang, 2007) was used to estimate lineage-specific rates of the nonsynonymous:synonymous substitution (dN/dS) ratio, ω. To detect whether positive selection had acted at some amino acid sites along particular lineages, a branch-site analysis was also performed by comparing the nearly neutral model (M1a) with Model A (MA), in which the foreground branch may have a proportion of sites under positive selection. All codeml analyses were implemented in EasyCodeML v1.0 (https://www.github.io/bioeasy/EasyCodeml).

Transcriptome sequencing

Samples of fresh leaf and root were collected from *D. zingiberensis*. Total RNA was isolated using the TRIzol reagent (Invitrogen, USA), followed by treatment with RNase-free DNase I (Promega, USA) according to the manufacturer' protocols. RNA quality was checked using an Agilent 2100 Bioanalyzer. Illumina RNA sequencing libraries were prepared for two samples and sequenced on a HiSeq 2500 system using a PE150 strategy following the manufacturer's instructions (Illumina, USA). After trimming based on quality scores with Btrim (Yong, 2011), the clean reads were aligned to the *D. zingiberensis* reference assembly using TopHat (Trapnell et al., 2014), and the mapped reads for each sample were assembled using Cufflinks (Trapnell et al., 2014). We used FPKM (fragments per kilobase of exon model per million mapped fragments) as the normalized gene expression level.

Comparison of transcriptomes among 11 plants

Except for *D. zingiberensis*, the raw RNA sequencing data for ten other plants were downloaded from the NCBI SRA database with the filters "Strategy = RNA-Seq AND Layout = Paired AND Tissue = Root or Tuber" (Supplemental Table 12). Gene FPKM values for each plant were calculated with TopHat and Cufflinks using a process similar to that described for *D. zingiberensis*. To make the gene expression among the 11 plants comparable, we first used OrthoMCL to identify one-to-one orthologous genes between *D. zingiberensis* and the other ten plants. Then, according to the median gene expression in *D. zingiberensis*, the gene expression of the other ten plants was normalized by dividing by the ratio of the median of gene expression in the target plant to the median of gene expression in *D. zingiberensis.*

Construction of the engineered yeast strain

We developed a stable cholesterol-producing strain as described in a previous study (Souza et al., 2011). First, the yeast native *ATF2* gene responsible for both steroid C3-hydroxy acetylation and prevention of further steroid metabolism was disrupted by CRISPR/Cas9 in strain BY-T3 (Supplemental Figure 6). Then, the dehydrocholesterol 7-reductase (*DHCR7*) and *DHCR24* coding sequences from *Danio rerio* were synthesized by GenScript (Nanjing, China) with codon optimization for *S. cerevisiae*. The homologous recombination-based DNA multi-segment assembly technology was used for multi-gene expression in yeast as described previously (Dai et al., 2013). The *DrDHCR7* gene was cloned into the pM2 plasmid under the control of the *PGK1* promoter and *ADH1* terminator, and *DrDHCR24* was cloned into the expression plasmid pYES2.0 with the *Gal1* promoter and the *CYC1* terminator. Subsequently, the *DrDHCR7* and *DrDHCR24* gene cassettes were amplified by PCR and mixed with an additional two DNA homologous arm fragments of the *Gal80* site. Finally, the mixed DNA was transformed into the *S. cerevisiae* δATF2 strain as described previously (Dai et al., 2019). The positive transformants were verified by PCR analysis, yielding the strain DG-Cho (Figure 3B). All the strains, primers, and plasmids used in this work are summarized in Supplemental Tables 11 and 12.

By analyzing the genome of *D. zingiberensis*, *DzinCYP90G6* and *DzinCYP94D144* were identified as cholesterol C-16,22-dihydroxylase and C-26 hydroxylase candidates, respectively. To identify the function of *DzinCYP90G6*, *VvCPR* (CPR from *Vitis vinifera*) with the *PGK1* promoter and *ADH1* terminator and *DzinCYP90G6* with the *TEF1* promoter and *CYC1* terminator were integrated into the *Gal7* site of strain DG-Cho, as described previously, to obtain strain DG001. Then, *DzinCYP94D144* with the *PGK1* promoter and *ADH1* terminator was integrated into the *ADH1* site of strain DG001. Strain DG002 was then developed to identify

the function of *DzinCYP94X*, and the resultant strain was able to convert cholesterol to diosgenin.

Fermentation

Engineered yeast strains were grown in SD medium containing 2% glucose and lacking leucine, uracil, tryptophan, and histidine, as appropriate (Dai et al., 2013). All strains were first inoculated into 15-ml culture tubes containing 2 ml medium and grown at 30 °C and 250 rpm to an optical density at 600 nm (OD_{600}) of approximately 2.0. Flasks (250 ml) containing 15 ml medium were then inoculated with the seed cultures to an OD_{600} of 0.05. Strains were grown at 30 °C and 250 rpm. After 30 h cultivation, cells were collected from fermentation culture via centrifugation, washed twice, resuspended in SD medium containing 2% galactose, and allowed to culture for 90 h. OD_{600} values were measured using a Shimadzu UV-2550 spectrophotometer.

GC-MS analysis of cholesterol and its hydroxyl products

Gas chromatography-mass spectrometry (GC-MS) analysis was performed with the following adaptations. Cells were collected from fermentation culture via centrifugation. The mixed solution (600 μl; acetone:methanol = 1:1) was added to the tube and crushed using a BeadBeater (BioSpec, USA) three times. The samples were then centrifuged at 10 000 *g* for 1 min, and 1 μl of supernatant was analyzed by GC-MS using an Agilent Technologies 5975C Inert xl MSD with Triple-Axis Detector equipped with an HP-5ms (30 m × 0.25 mm × 0.5 μm) GC column. Compound separation was achieved with an injector temperature of 300 °C and a 36-min temperature gradient program for GC separation starting at 240 °C for 5 min, followed by heating the column to 300 °C at 10 °C min^{-1} and a final constant hold at 300 °C for 25 min. Mass detection was achieved with electric ionization using SIM scan mode with diagnostic ions monitored as follows: m/z 69, m/z 139, m/z 282, and m/z 414. Cholesterol and diosgenin samples purchased from Yuanye (Shanghai, China) were used as standards.

GC-MS analysis of diosgenin in monocots

To detect diosgenin, the frozen plant materials were ground in a mortar under liquid nitrogen. The resulting powder was dissolved in methanol/acetone solution (v/v = 1:1), shaken vigorously using a vortex oscillator, and ultra-sonicated for 30 min. The cell debris was removed by centrifugation at 13 000 *g* for 30 min. The supernatant was used for GC-MS analysis according to the GC-MS analysis method described above for cholesterol, with the only difference being the use of more precise equipment: an Agilent 7890A GC/7200 Accurate-MQ-TOF MS system coupled with an HP-5ms (30 m × 0.25 mm× 0.5 μm) GC column.

Correlation analysis of gene expression and CpG density

To investigate the relationship between gene expression and CpG density (including CG dinucleotide and CHG trinucleotide, where H indicates any nucleotide except guanine) at the promoters of genes related to diosgenin and starch biosynthesis in *D. zingiberensis* and *D. rotundata*, the CpG density from 1000 bp upstream of the transcription start site to 100 bp downstream was calculated using a window size of 50 nt and a walking step of 10 nt. To eliminate the disturbance caused by the overall difference in GC content of the two genomes, the CpG density in the focused region of each gene was normalized by the average CpG density of all genes in the corresponding genome. The average CpG densities of genes involved in diosgenin and starch biosynthesis were calculated for *D. zingiberensis* and *D. rotundata*, respectively. Student's *t*-test was used to test the statistical significance of the difference in CpG density between the two species.

DATA AND CODE AVAILABILITY

The *D. zingiberensis* genome sequence has been deposited at NCBI under BioProject ID PRJNA541739 and accession number VCDL00000000.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Plant Communications Online*.

FUNDING

This work was supported by grants from the National Key R&D Program of China (no. 2019YFA0905700 and 2019YFA0905300), the Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (TSBICIP-KJGG-002), the Key Research Program of the Chinese Academy of Sciences (KFZD-SW-215), the Tianjin Science Fund for Distinguished Young Scholars (18JCJQJC48300), the National Science and Technology Major Project (2018ZX09711001-006-003), the Major Science and Technique Programs in Yunnan Province (2019ZF011), and the National Science Fund for Excellent Young Scholars (31922047).

AUTHOR CONTRIBUTIONS

X. Zhang and H. Jiang designed the study. J. Chen and X. Liu prepared materials for genomic and RNA sequencing analysis. J. Cheng performed the genomic analysis and evolutionary analysis. J. Chen constructed the cholesterol-producing chassis and identified two diosgenin biosynthetic genes. X. Liu performed the GC-MS analysis of diosgenin in monocots. X. Li constructed the phylogenetic tree of 15 plants and performed the molecular evolution analyses of the CYP90 family. W. Zhang and J. Cai performed the correlation analysis of gene expression and CpG density. H. Jiang, J. Cheng, and J. Chen wrote the manuscript. All authors discussed the results and commented on the manuscript.

ACKNOWLEDGMENTS

This work has been included in patent applications by the Tianjin Institute of Industrial Biotechnology.

Received: December 13, 2019
Revised: May 25, 2020
Accepted: May 29, 2020
Published: June 2, 2020

REFERENCES

Alda, L., Gogoaşă, I., Bordean, D.-M., Gergen, I., Alda, S., Moldovan, C., and Niţă, L. (2009). Lycopene content of tomatoes and tomato products. J. Agroaliment. Proc. Technol. **15**:540–542.

Bai, Y., Zhang, L., Jin, W., Wei, M., Zhou, P., Zheng, G., Niu, L., Nie, L., Zhang, Y., and Wang, H. (2015). In situ high-valued utilization and transformation of sugars from *Dioscorea zingiberensis* C.H. Wright for clean production of diosgenin. Bioresour. Technol. **196**:642–647.

Balandrin, M.F., Klocke, J.A., Wurtele, E.S., and Bollinger, W.H. (1985). Natural plant chemicals: sources of industrial and medicinal materials. Science **228**:1154–1160.

Christ, B., Xu, C., Xu, M., Li, F.-S., Wada, N., Mitchell, A.J., Han, X.-L., Wen, M.-L., Fujita, M., and Weng, J.-K. (2019). Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. Nat. Commun. **10**:3206.

Dai, Z., Liu, Y., Sun, Z., Wang, D., Qu, G., Ma, X., Fan, F., Zhang, L., Li, S., and Zhang, X. (2019). Identification of a novel cytochrome P450 enzyme that catalyzes the C-2α hydroxylation of pentacyclic triterpenoids and its application in yeast cell factories. Metab. Eng. **51**:70–78.

Dai, Z., Liu, Y., Zhang, X., Shi, M., Wang, B., Wang, D., Huang, L., and Zhang, X. (2013). Metabolic engineering of *Saccharomyces cerevisiae* for production of ginsenosides. Metab. Eng. **20**:146–156.

De Luca, V., Salim, V., Atsumi, S.M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. Science **336**:1658.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. **25**:1010.

Dixon, R.A. (2001). Natural products and plant disease resistance. Nature **411**:843–847.

Dudareva, N., Negre, F., Nagegowda, D., and Orlova, I. (2006). Plant volatiles: recent advances and future perspectives. Crit. Rev. Plant Sci. **25**:417–440.

Fernandes, P., Cruz, A., Angelova, B., Pinheiro, H.M., and Cabral, J.M.S. (2003). Microbial conversion of steroid compounds: recent developments. Enzyme Microb. Technol. **32**:688–705.

Florence Ménard, S., Cathy, C., Coralie, V., Amélie, M., Stéphane, B., Catherine, D., Sophie, B., Agnès, G., Eric, T., and Patricia, C. (2003). Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nat. Biotechnol. **21**:143–149.

Govaerts, R., Wilkin, P., and Saunders, R. (2007). World Checklist of Dioscorales: Yams and Their Allies (Royal Botanic Gardens Kew), pp. 1–65.

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., Teodor, R., Lu, Y., Bowser, T.A., Graham, I.A., et al. (2018). The opium poppy genome and morphinan production. Science **362**:343.

Guo, N., Cheng, F., Wu, J., Liu, B., Zheng, S., Liang, J., and Wang, X. (2014). Anthocyanin biosynthetic genes in *Brassica rapa*. BMC genomics **15**:426.

He, Z., Tian, Y., Zhang, X., Bing, B., Zhang, L., Wang, H., and Zhao, W. (2012). Anti-tumour and immunomodulating activities of diosgenin, a naturally occurring steroidal saponin. Nat. Product Res. **26**:2243–2246.

Huang, H., Shanlin, G., Lanlan, C., and Xiaoke, J. (2008). In vitro induction and identification of autotetraploids of *Dioscorea zingiberensis*. Vitro Cell Dev. Biol. Plant **44**:448–455.

Jing, C., Xin, L., Kevin, V., Sebastian, P., Wen-Chieh, T., Ke-Wei, L., Li-Jun, C., Ying, H., Qing, X., and Chao, B. (2015). Corrigendum: the genome sequence of the orchid *Phalaenopsis equestris*. Nat. Genet. **47**:65–72.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**:772–780.

Kevin, V., Steven, M., and Yves, V.D.P. (2014). Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. Lond. **369**:5042–5050.

Li, M., Zhang, D., Gao, Q., Luo, Y., Zhang, H., Ma, B., Chen, C., Whibley, A., Zhang, Y.e., Cao, Y., et al. (2019). Genome structure and evolution of *Antirrhinum majus* L. Nat. Plants **5**:174–183.

Liu, X., Cheng, J., Zhang, G., Ding, W., Duan, L., Yang, J., Kui, L., Cheng, X., Ruan, J., and Fan, W. (2018). Engineering yeast for the production of breviscapine by genomic analysis and synthetic biology approaches. Nat. Commun. **9**:448.

Maeda, H.A. (2019). Evolutionary diversification of primary metabolism and its contribution to plant chemical diversity. Front. Plant Sci. **10**. https://doi.org/10.3389/fpls.2019.00881.

Mafalda, J., Martins, A.P.J., Eugenia, G., and Samuel, S. (2016). Diosgenin: recent highlights on pharmacology and analytical methodology. J. Anal. Methods Chem. **2016**:1–16.

Mignouna, H.D., Abang, M.M., Asiedu, R., and Geeta, R. (2009). True yams (Dioscorea): a biological and evolutionary link between eudicots and grasses. Cold Spring Harbor Protoc. https://doi.org/10.1101/pdb.emo136.

Moses, T., Papadopoulou, K.K., and Osbourn, A. (2014). Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives. Crit. Rev. Biochem. Mol. Biol. **49**:439.

Nadeem, M., Rikhari, H., Kumar, A., Palni, L., and Nandi, S. (2002). Taxol content in the bark of Himalayan Yew in relation to tree age and sex. Phytochemistry **60**:627–631.

Osbourn, A.E., and Lanzotti, V. (2009). Plant-Derived Natural Products (Springer), pp. 1–32.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061–1067.

Pichersky, E., and Raguso, R.A. (2016). Why do plants produce so many terpenoid compounds? New Phytol. **220**:692–702.

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., and Qi, J. (2018). Wide-spread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol. Plant **11**:414–428.

Renne, P.R., Deino, A.L., Hilgen, F.J., Kuiper, K.F., Mark, D.F., Mitchell, W.S., Morgan, L.E., Roland, M., and Jan, S. (2013). Time scales of critical events around the Cretaceous-Paleogene boundary. Science **339**:684–687.

Satomi, F., Toshiyuki, O., Bunta, W., Takao, Y., Suguru, T., Shozo, F., Shigeo, Y., Kanzo, S., and Masaharu, M. (2010). Arabidopsis CYP90B1 catalyses the early C-22 hydroxylation of C27, C28 and C29 sterols. Plant J. **45**:765–774.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Sonawane, P.D., Pollier, J., Panda, S., Szymanski, J., Massalha, H., Yona, M., Unger, T., Malitsky, S., Arendt, P., and Pauwels, L. (2016). Plant cholesterol biosynthetic pathway overlaps with phytosterol metabolism. Nat. Plants **3**:16205.

Souza, C.M., Schwabe, T.M.E., Pichler, H., Ploier, B., Leitner, E., Guan, X.L., Wenk, M.R., Riezman, I., and Riezman, H. (2011). A stable yeast strain efficiently producing cholesterol instead of ergosterol is functional for tryptophan uptake, but not weak organic acid resistance. Metab. Eng. **13**:555–569.

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

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**:W609.

Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., Yoshida, K., Uemura, A., Oikawa, K., Abe, A., et al. (2017). Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. BMC Biol. **15**:86.

Tang, W., and Eisenbrand, G. (1993). Chinese Drugs of Plant Origin: Chemistry, Pharmacology, and Use in Traditional and Modern Medicine, 32 (Springer Science & Business Media), pp. 1081–1090.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2014). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7**:562–578.

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., Yang, M., He, L., Deng, T., Escalante, F.J., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. Proc. Natl. Acad. Sci. U S A **114**:E9413–E9422.

Wang, S., Alseekh, S., Fernie, A.R., and Luo, J. (2019). The structure and function of major plant metabolite modifications. Mol. Plant **12**:899–919.

Wang, Y., Zhang, Y., Zhu, Z., Zhu, S., Li, Y., Li, M., and Yu, B. (2007). Exploration of the correlation between the structure, hemolytic activity, and cytotoxicity of steroid saponins. Bioorg. Med. Chem. **15**:2528–2532.

Weng, J.K., and Noel, J.P. (2012). The rise of chemodiversity in plants. Science **336**:1667–1670.

Yang, H., Yin, H.-W., Wang, X.-W., Li, Z.-H., Shen, Y.-P., and Jia, X.-B. (2015). In situ pressurized biphase acid hydrolysis, a promising approach to produce bioactive diosgenin from the tubers of *Dioscorea zingiberensis*. Pharmacogn Mag. **11**:636–642.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586.

Yong, K. (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics **98**:152–153.

Yosef Al Jasem, M.K., Ahmed, T., and Thiemann, T. (2014). Preparation of steroidal hormones with an emphasis on transformations of phytosterols and cholesterol—a review. Mediterr. J. Chem. **3**:796–830.

Zhixi, T., Qian, Q., Qiaoquan, L., Meixian, Y., Xinfang, L., Changjie, Y., Guifu, L., Zhenyu, G., Shuzhu, T., and Dali, Z. (2009). Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. Proc. Natl. Acad. Sci. U S A **106**:21760–21765.