The Reference Genome Sequence of Scutellaria baicalensis Provides Insights into the Evolution of Wogonin Biosynthesis

Molecular Plant
Research Article

CellPress
PARTNER JOURNAL

The Reference Genome Sequence of *Scutellaria baicalensis* Provides Insights into the Evolution of Wogonin Biosynthesis

Qing Zhao[1,2,8], Jun Yang[1,2,8], Meng-Ying Cui[1], Jie Liu[1], Yumin Fang[1], Mengxiao Yan[1,2], Wenqing Qiu[3], Huiwen Shang[4], Zhicheng Xu[4], Reheman Yidiresi[4], Jing-Ke Weng[5,6], Tomáš Pluskal[5], Marielle Vigouroux[7], Burkhard Steuernagel[7], Yukun Wei[1], Lei Yang[1], Yonghong Hu[1], Xiao-Ya Chen[1,2] and Cathie Martin[1,7,*]

[1]Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences, Shanghai, China

[2]State Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

[3]Key Laboratory of Metabolism and Molecular Medicine, Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Fudan University, Shanghai, China

[4]Novogene Bioinformatics Institute, Beijing, China

[5]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142, USA

[6]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[7]John Innes Centre, Norwich NR4 7UH, UK

[8]These authors contributed equally to this article.

***Correspondence: Cathie Martin (cathie.martin@jic.ac.uk)**

https://doi.org/10.1016/j.molp.2019.04.002

ABSTRACT

***Scutellaria baicalensis* Georgi is important in Chinese traditional medicine where preparations of dried roots, "Huang Qin," are used for liver and lung complaints and as complementary cancer treatments. We report a high-quality reference genome sequence for *S. baicalensis* where 93% of the 408.14-Mb genome has been assembled into nine pseudochromosomes with a super-N50 of 33.2 Mb. Comparison of this sequence with those of closely related species in the order Lamiales, *Sesamum indicum* and *Salvia splendens*, revealed that a specialized metabolic pathway for the synthesis of 4′-deoxyflavone bioactives evolved in the genus *Scutellaria*. We found that the gene encoding a specific cinnamate coenzyme A ligase likely obtained its new function following recent mutations, and that four genes encoding enzymes in the 4′-deoxyflavone pathway are present as tandem repeats in the genome of *S. baicalensis*. Further analyses revealed that gene duplications, segmental duplication, gene amplification, and point mutations coupled to gene neo- and subfunctionalizations were involved in the evolution of 4′-deoxyflavone synthesis in the genus *Scutellaria*. Our study not only provides significant insight into the evolution of specific flavone biosynthetic pathways in the mint family, Lamiaceae, but also will facilitate the development of tools for enhancing bioactive productivity by metabolic engineering in microbes or by molecular breeding in plants. The reference genome of *S. baicalensis* is also useful for improving the genome assemblies for other members of the mint family and offers an important foundation for decoding the synthetic pathways of bioactive compounds in medicinal plants.**

Key words: genome, skullcap, 4′-deoxyflavone, traditional Chinese medicine, Huang Qin, evolution, specialized metabolism

Zhao Q., Yang J., Cui M.-Y., Liu J., Fang Y., Yan M., Qiu W., Shang H., Xu Z., Yidiresi R., Weng J.-K., Pluskal T., Vigouroux M., Steuernagel B., Wei Y., Yang L., Hu Y., Chen X.-Y., and Martin C. (2019). The Reference Genome Sequence of *Scutellaria baicalensis* Provides Insights into the Evolution of Wogonin Biosynthesis. Mol. Plant. **12**, 935–950.

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

INTRODUCTION

Scutellaria baicalensis Georgi, or Chinese skullcap, is a well-known medicinal plant that is cultivated worldwide for its therapeutic properties (Shang et al., 2010). The dried root of *S. baicalensis* has been used as a traditional medicine for more than 2000 years in China, where it is called Huang Qin (黄芩) (Li, 2012; Zhao et al., 2016a). Huang Qin is used for treatment of bitter, cold, liver, and lung problems as recorded by *Shennong Bencao Jing* (The Divine Farmer's Materia Medica) written between 200 and 250 AD (Ma, 2013). Recent studies have reported the pharmacological activities of *Scutellaria* root preparations, particularly those of novel flavonoids (Li-Weber, 2009; Shang et al., 2010; Qiao et al., 2016; Tu et al., 2016). The bioactivities of the root flavonoids of *S. baicalensis* include antibacterial, antiviral, antioxidant, anticancer, hepatoprotective, and neuroprotective properties (Wen, 2007; Gao et al., 2011; Yang et al., 2012). Despite the commercial interest and increasing demand for *Scutellaria*, improvements through breeding have been very limited. The absence of genome information has limited the understanding of how its flavonoid bioactives are made and have limited any improvement in productivity through genetic selection. Understanding the genes responsible for biosynthesis of the various flavonoids made in *S. baicalensis* and their regulation will lay a foundation for molecular breeding for improved, sustainable production.

Two pathways operate in *S. baicalensis* for the synthesis of flavones. In the aerial parts of the plant, expression of the classic flavone biosynthetic pathway genes encoding phenylalanine ammonia lyase (SbPAL), cinnamate 4-hydroxylase (SbC4H), 4-coumarate coenzyme A (CoA) ligase (SbCLL1), chalcone synthase (SbCHS-1), chalcone isomerase (SbCHI), and flavone synthase II (SbFNSII1), leads to the production of the 4′-hydroxyflavone apigenin, which is hydroxylated and glycosylated to form scutellarein and scutellarin, respectively (Zhao et al., 2016b). In roots, a new specialized pathway producing root-specific 4′-deoxyflavones has evolved relatively recently (Zhao et al., 2016b). The root-specific pathway recruits a cinnamate-CoA ligase (SbCLL-7) from fatty acid metabolism to form cinnamoyl-CoA from cinnamate produced by SbPAL. Cinnamoyl-CoA is then condensed with malonyl-CoA by SbCHS-2, a CHS isoform that evolved from SbCHS-1, to specifically produce pinocembrin chalcone. The chalcone is isomerized by the same CHI that operates in the classic flavone pathway in aerial parts to form pinocembrin, a flavanone lacking a 4′-OH group (Figure 1). Chrysin is formed from pinocembrin by a specialized isoform of flavone synthase, named FNSII-2 (Zhao et al., 2016a). Chrysin is the first root-specific flavone (RSF) and may be decorated by different types of flavone hydroxylases, methyltransferases, and glycosyltransferases to form the other RSFs found in the roots of *S. baicalensis* as shown in Figure 1 (Zhao et al., 2018). Baicalein, wogonin, and their glycosides baicalin and wogonoside, are the major RSFs with explicit pharmacological activities in extracts of the roots of *S. baicalensis* (Kovács et al., 2004; Islam et al., 2011; Qiao et al., 2016). To date, we have elucidated the entire biosynthetic pathway for baicalein and norwogonin (Zhao et al., 2018). However, the final step for biosynthesis of wogonin from norwogonin, which is carried out by an 8-*O*-methyltransferase, remains to be elucidated.

Whole-genome sequencing has become a practical strategy to identify metabolic pathways for natural product biosynthesis (Liu et al., 2017; Mochida et al., 2017), and genome sequences of members of the order Lamiales such as *Salvia miltiorrhiza* (Zhang et al., 2015; Xu et al., 2016), *Salvia splendens* (Dong et al., 2018), and *Sesamum indicum* (Zhang et al., 2013) make feasible the comparative analysis to investigate how RSFs are produced in *Scutellaria*.

Here we report a reference genome sequence of *S. baicalensis* obtained using a combination of Illumina and PacBio data, which was assembled using information from 10x Genomics and Hi-C technologies. In total, 386.63 Mb of the 408.14-Mb genome were assembled, and the sequences were sorted into nine pseudochromosomes with a super-N50 of 33.9 Mb. Comparative genomic analysis was performed with the published genomes of *S. miltiorrhiza*, *S. splendens*, and *S. indicum*. The evolutionary path for the biosynthesis of RSFs appears to have arisen by specific recruitment of a gene encoding a CoA ligase (Figure 1; *SbCLL-7*) as well as several tandem gene replications (Figure 1: *SbCHS-2*, *SbFNSII-2*, *SbF8H*). The *O*-methyltransferases (OMTs) responsible for wogonin synthesis were identified by screening the genome and transcriptome sequences and were confirmed by *in vivo* assays in yeast as well as by RNAi experiments in hairy roots of *S. baicalensis*.

RESULTS

Genome Sequencing, Assembly, and Annotation

The DNA for genome sequencing of *S. baicalensis* came from a single plant maintained in Shanghai Chenshan Botanical Garden. DNA was extracted and sequenced using Illumina and PacBio sequencing strategies. We obtained 48.02 Gb of PacBio reads, amounting to 117.66× coverage of the 408.14-Mb genome, a size estimated by k-mer distribution analysis (Supplemental Figure 1A and Supplemental Table 1). We experimentally determined the genome size to be 392 Mb using flow cytometry, which is close to the value given by the k-mer method (Supplemental Figure 1B). After interactive error correction among the PacBio reads, assembly was carried out using FALCON to obtain primary contigs. To avoid problems due to heterozygosity from outcrossing, we phased the contigs using FALCON-Unzip and polished the updated primary contigs and haplotigs with Quiver. The final contigs were error corrected with 67.96 Gb (166.51×) of short reads obtained from Illumina HiSeq X Ten sequencing (Supplemental Table 1). The consensus sequences were further assembled using the reference of 86.72 Gb (212.485× coverage) from 10x Genomics sequencing (Supplemental Table 1). All the contigs were extended using FragScaff to generate an assembly with a total scaffold length of 386.6 Mb (94.73% of the genome) and an N50 of 1.33 Mb (Supplemental Table 2). To facilitate genome annotation and to obtain expression profiles of the *S. baicalensis* genes, we sequenced RNA samples from flowers, flower buds, leaves, stems, roots, and jasmonic acid (JA)-treated roots from multiple lines of *S. baicalensis* (Zhao et al., 2016a, 2018). RNA samples from each tissue were extracted in triplicate and sequenced using the HiSeq 2000 platform.

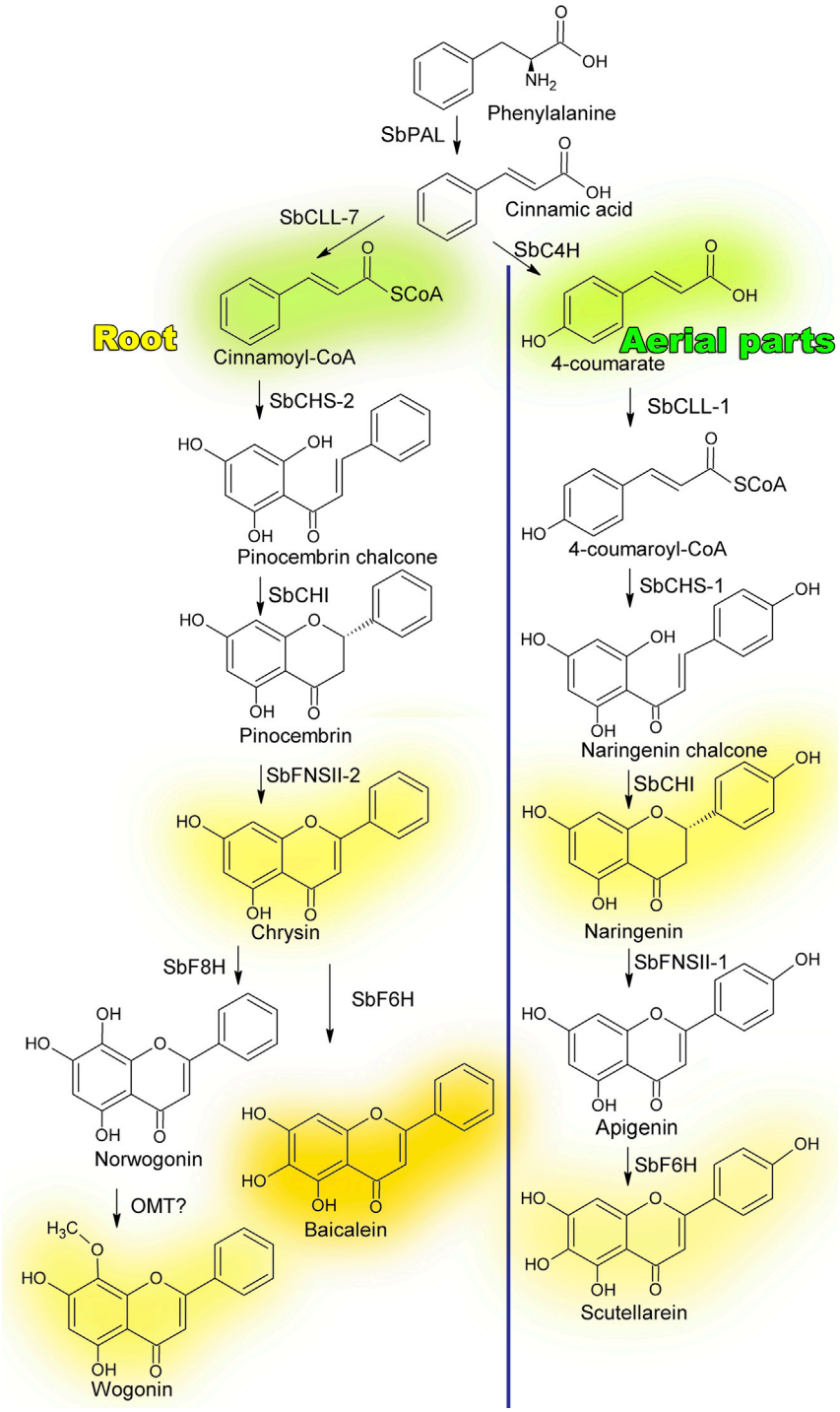

Figure 1. Two Pathways for Biosynthesis of Flavones in *S. baicalensis*.
In aerial tissues, the classic flavone pathway makes scutellarein, while in roots, root-specific flavones, baicalein, and wogonin are produced by a newly evolved pathway. Enzymes are: phenylalanine ammonia lyase (SbPAL), cinnamate 4-hydroxylase (SbC4H), cinnamate-CoA ligase (SbCLL-7), 4-coumarate CoA ligase (SbCLL-1), chalcone synthase (SbCHS-1), pinocembrin-chalcone synthase (SbCHS-2), chalcone isomerase (SbCHI), flavone synthase II (SbFNSII), flavone 6-hydroxylase (SbF6H), flavone 8-hydroxylase (SbF8H), and 8-*O*-methyl transferase (OMT).

An Hi-C (*in vivo* fixation of chromosomes) library was then employed to refine the first version of the reference genome. This method sorted 475 of the 578 scaffolds into 114 super-scaffolds, accounting for 98.04% of the original 386.63-Mb assembly (Supplemental Tables 2 and 3). All the super-scaffolds could be placed in one of nine groups (Supplemental Figure 2). The super-scaffold N50 reached 33.2 Mb, with the longest super-scaffold being 87.96 Mb (Supplemental Tables 3 and 4). The number of groups, hereafter referred to as pseudochromosomes, corresponded well to the number of chromosomes reported previously ($1n$ = 9, $2n$ = 18) (Cheng et al., 2010). The genome of *S. baicalensis* has a GC content of 34.24%, with N comprising 0.6% (Supplemental Table 5). SNP calling based on the genome sequence revealed a heterozygosity rate of 0.31% (Supplemental Table 6).

To test the coverage of the genome, we mapped the short reads generated from Illumina sequencing, and 96.5% of these reads could be mapped to the scaffolds with 99.78% overall coverage (Supplemental Table 7). Expressed sequence tag (EST) sequences generated from transcriptome sequencing were also mapped to the assembly, and 86.12% of ESTs had more than 90% of their sequences in one scaffold while 97.51% of ESTs had more than 50% of their sequences in one scaffold (Supplemental Table 8). CEGMA (Core Eukaryotic Genes Mapping Approach) and BUSCO (Benchmarking Universal Single-Copy Orthologs) evaluations of the genome sequence indicated 96.37% and 94% completeness, respectively (Supplemental Tables 9 and 10). All analyses suggested a good quality of assembly.

A pipeline combining *de novo* predictions, homology-based predictions, and RNA-sequencing data was used to construct gene models for the *S. baicalensis* genome. A total of 28 930 genes were annotated this way, with an average length of 2980 bp and an average coding sequence length of 1122 bp (Supplemental Tables 3 and 11), of which 23 027 genes (79.6%) were supported by RNA-sequencing data. A total of 20 234 genes (69.9%) were supported by all three methods (RNA sequencing, *de novo* predictions, and homology), and these genes were annotated with high confidence. The resulting protein models were then compared with protein sequences in four protein databases; NR, SwissProt, KEGG, and InterPro. We found that 28 524 (98.6%) gene products could be annotated by at least one of the databases. Genes were named according to the nomenclature used for *Arabidopsis*

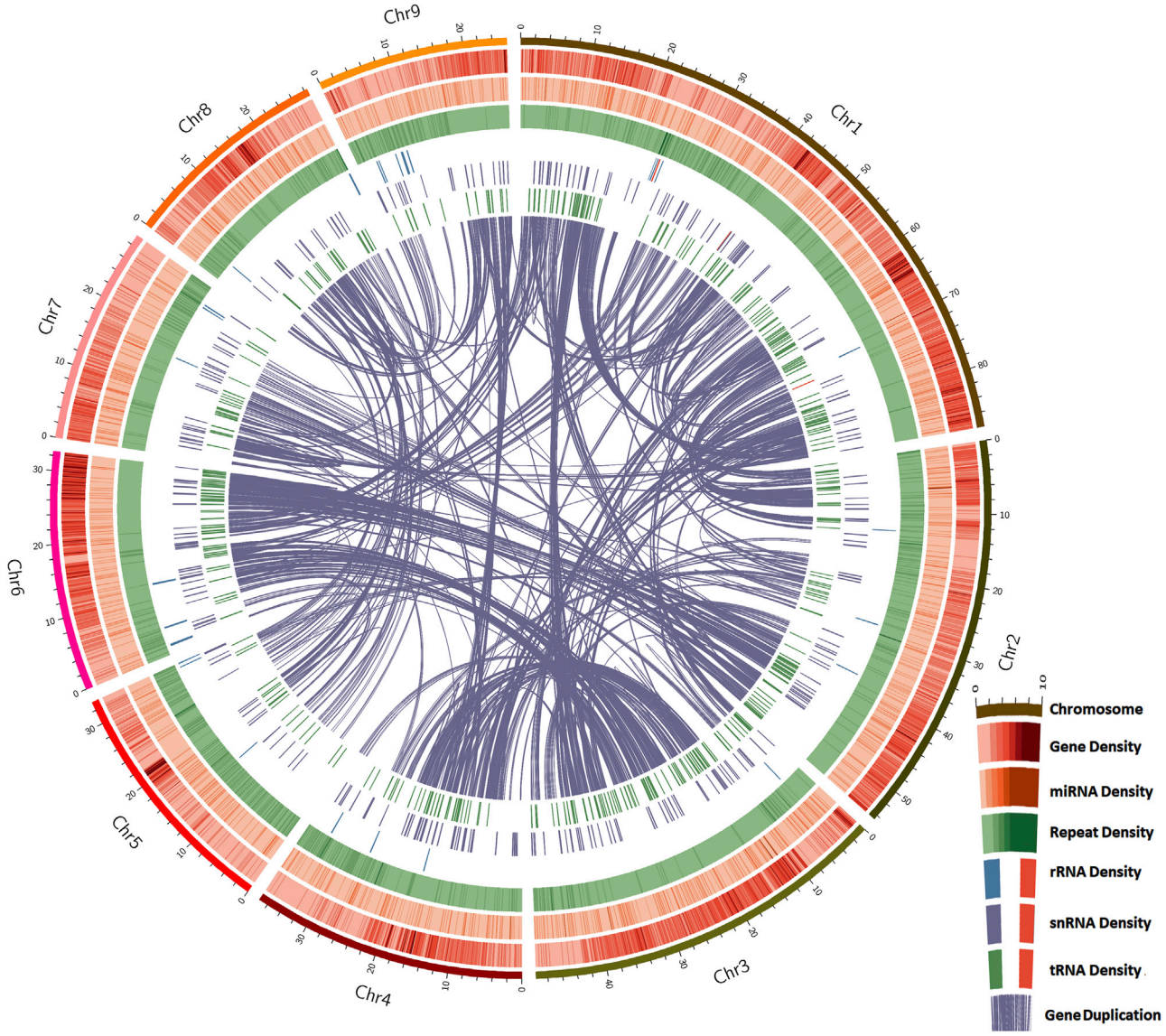

Figure 2. Overview of *S. baicalensis* Draft Genome Assembly.
The outer lines represent pseudochromosomes. The colored bands summarize the density of genes (red), miRNAs (orange), repeats (green), rRNAs (blue), snRNAs (purple), and tRNAs (dark green). All detected gene duplications are indicated with links inside the circles. Scales show chromosomes in a 0- to 10-kb window; gene density in a 100-kb window (0–29, which means the density indicated by the color gradient starts from 0 and goes to 29 genes per 100 kb DNA); miRNA density in a 100-kb window (0–5); repeat density in a 100-kb window (0–2.67); rRNA density in a 100-kb window (0–274); snRNA density in a 100-kb window (0–24); tRNA density in a 100-kb window (0–22); detected gene duplication links (3516). The red bar in the rRNA, snRNA, and tRNA keys indicates the maximum density of copies on the scale.

(*Arabidopsis* Genome Initiative, 2000) to indicate the relative positions of genes on the pseudochromosomes.

The assembled draft *S. baicalensis* genome contains 55.15% repetitive sequences. Tandem duplications (small satellites and microsatellites) and interspersed repeats accounted for 1.2% and 53.95% of the genome, respectively. Long terminal repeats (LTRs) of retroelements were the most abundant interspersed repeat, occupying 34.4% of the genome, followed by DNA transposable elements at 15.4% (Supplemental Table 12). Genes annotated as encoding non-coding RNAs (ncRNAs) in the current genome included 1218 microRNAs (miRNAs), 517 transfer RNAs (tRNAs), 1846 ribosomal RNAs (rRNAs), and 512 small nuclear RNAs (snRNAs) (Supplemental Table 13). An overview of the genes, repeats, non-coding RNA densities, and all detected segmental duplications is presented in Figure 2.

Comparative Genomic Analysis

We compared our assembly for *S. baicalensis* with 10 other sequenced genomes from four eurosid species (*Arabidopsis thaliana*, *Populus trichocarpa*, *Glycine max*, *Vitis vinifera*), four euasterid species (*Solanum lycopersicum*, *S. indicum*, *S. splendens*, *S. miltiorrhiza*), a monocot, *Oryza sativa*, and *Amborella trichopoda*, which, as a sister group to all other flowering plants, represents a species at the base of the angiosperms. Based on

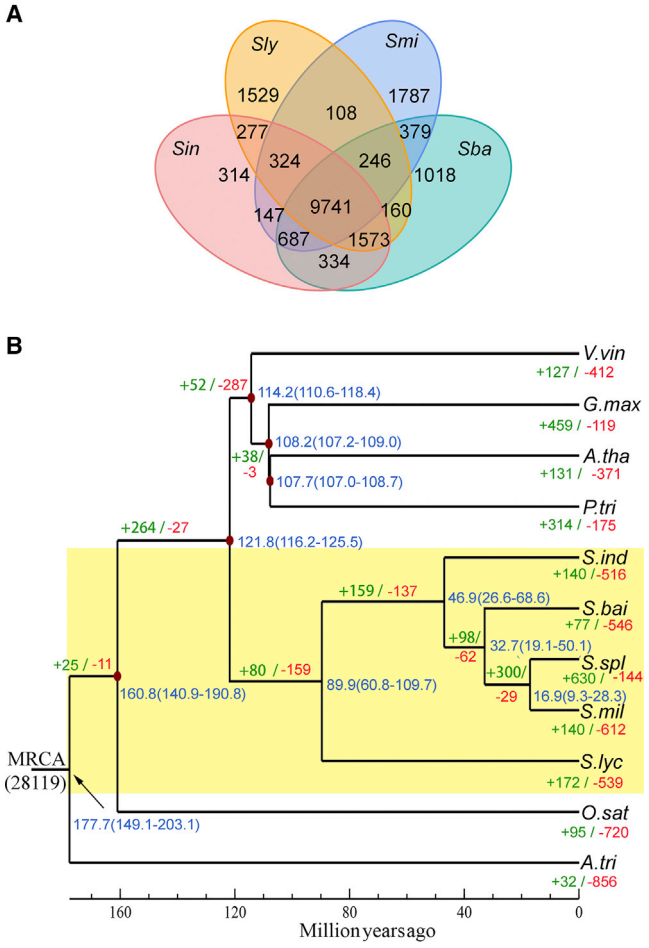

Figure 3. Comparative Genomic Analysis.

(A) Comparison of the number of gene families in *S. baicalensis* with that in other euasterid species. *Sin* refers to *S. indicum*, *Sly* refers to *S. lycopersicum*, *Smi* refers to *S. miltiorrhiza*, and *Sba* refers to *S. baicalensis*.

(B) Phylogenetic analysis and divergence time estimations among 11 plant species. The tree was constructed based on 299 single-copy truly orthologous genes. Divergence times (Mya) are indicated by the blue numbers beside the branch nodes. The number of gene-family contraction and expansion events is indicated by green and red numbers (respectively) below each species name. *V.vin*, *Vitis vinifera*; *G.max*, *Glycine max*; *A.tha*, *Arabidopsis thaliana*; *P.tri*, *Populus trichocarpa*; *S.ind*, *Sesamum indicum*; *S.bai*, *Scutellaria baicalensis*; *S.spl*, *Salvia splendens*; *S.mil*, *Salvia miltiorrhiza*; *S.lyc*, *Solanum lycopersicum*, *O.sat*, *Oryza sativa*; *A.tri*, *Amborella trichopoda*.

analysis of gene family clustering, we identified 28 133 gene families, of which 6811 were shared by all 11 species, and 299 of these shared families were single-copy gene families (Supplemental Figure 3). We compared the gene numbers among the four euasterid species. As shown in Figure 3A, 9741 gene families were shared by *S. baicalensis*, *S. indicum*, *S. lycopersicum*, and *S. miltiorrhiza*, and 1018 gene families were specific to *S. baicalensis*. *S. baicalensis* had fewer specific gene families than the other two members of the Lamiaceae; there were 1529 specific gene families in *S. lycopersicum* and 1787 in *S. miltiorrhiza*. *S. indicum* had only 314 specific gene families. These data support the conclusion of Xu et al. (2016) that genes identified by transcriptome analyses as involved in iridoid and monoterpenoid pathways are absent or greatly reduced in number in *S. baicalensis*, despite being present in other members of the order Lamiaceae.

The expanded and contracted gene families of the 11 plant species were compared with their most recent common ancestor (MRCA); 77 gene families were expanded and 546 gene families were contracted in *S. baicalensis*. Gene Ontology (GO) studies based on the 77 expanded gene families showed enrichment of genes encoding "UDP-glucosyltransferase activity" and "transferring hexosyl groups," suggesting the importance of glucodecoration of metabolites in *S. baicalensis*, and indeed most of the flavones found in *S. baicalensis* can be glycosylated.

Single-copy orthologous genes (299) were retrieved from the 11 species and alignments were undertaken based on these sequences. We combined all the alignments to produce a superalignment matrix, which was used to construct a phylogenetic tree. The branching order in the tree was consistent with a previously proposed phylogenetic ordering: *S. lycopersicum* (Solanaceae) diverged from Lamiales approximately 89.9 million years ago (Mya), and *S. indicum* diverged from the Lamiaceae about 46.9 Mya, followed by divergence of *S. baicalensis* from *Salvia* spp. about 32.7 Mya (Figure 3B) (Xu et al., 2016; Mint Evolutionary Genomics Consortium, 2018).

Screening for Genes Encoding Enzymes of the Flavone Biosynthetic Pathways

RSFs without a 4′-OH on the B-ring are the major bioactive compounds found in *S. baicalensis*. Our previous studies showed that a specialized flavone pathway probably evolved after divergence of the Lamiaceae (Zhao et al., 2016b, 2018). To find clues about the evolution of the RSF pathway, we screened the assembly for genes encoding enzymes of the classic flavone pathway and the RSF pathway (Figure 1 and Supplemental Table 14), and then compared the pathway genes in *S. baicalensis* with those in *S. miltiorrhiza* and *S. splendens* from the family Lamiaceae and with those in *S. indicum*, which belongs to the order Lamiales like *Salvia* and *Scutellaria*, but is in the family Pedaliaceae.

First analyzed were the *SbCLL-1* (locus ID Sb02g19320) and *SbCLL-7* (locus ID Sb09g15340) genes, which encode 4-coumarate-CoA ligase and cinnamate-CoA ligase, the first committed enzymes in the synthesis of scutellarein and RSFs, respectively. The regions of DNA where *SbCLL-1* and *SbCLL-7* are located are well conserved between *S. miltiorrhiza*, *S. splendens*, and *S. indicum* (Supplemental Figure 4A and 4B). The three other genomes have *SbCLL-1* and *SbCLL-7* homologs in synteny with *SbCLL-1* and *SbCLL-7* in *S. baicalensis*. *SbCLL-7* encodes a CoA ligase specific for cinnamate as its accepter and appears to have been recruited from a family of CoA ligases involved in fatty acid biosynthesis (Zhao et al., 2016b). We isolated cDNAs for the *CLL* genes in the regions of synteny from *S. miltiorrhiza*, *S. splendens*, and *S. indicum* and named them *SmCLL-7*, *SsCLL-7*, and *SiCLL-7*, respectively. The three *CLL-7* cDNAs, together with *SbCLL-7* cDNA (Zhao et al., 2016b) were expressed in *Escherichia coli* and the proteins were purified (Supplemental Figure 5A). The enzymes were assayed with three 4CL substrates: cinnamic acid, 4-coumaric acid, and

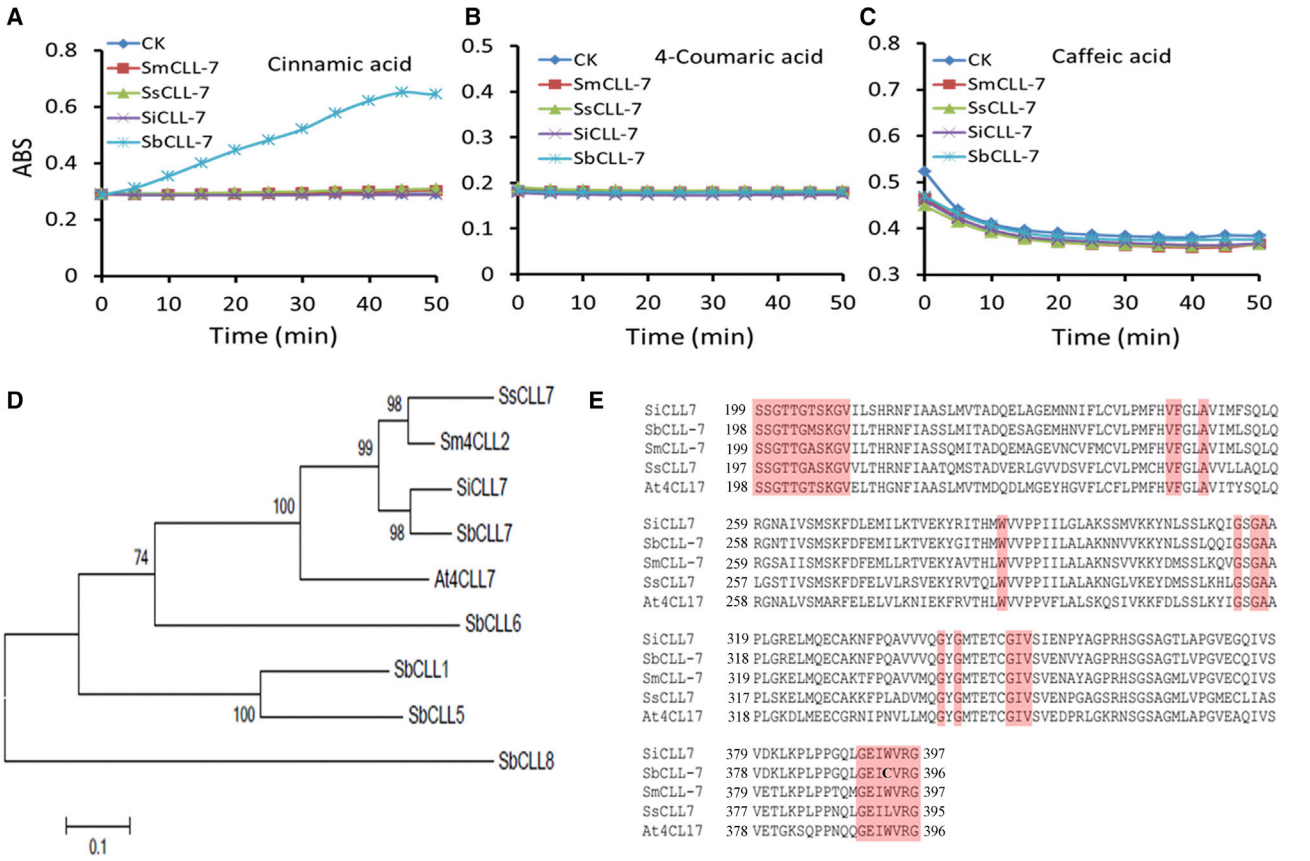

Figure 4. A Specific SbCLL-7 from *S. baicalensis*.

(A–C) Enzyme assays of CLL-7 proteins against cinnamic acid **(A)**, 4-coumaric acid **(B)**, and caffeic acid **(C)**. SbCLL-7, SmCLL-7, SsCLL-7, and SiCLL-7 were assayed against cinnamic acid, 4-coumaric acid, and caffeic acid. Absorbance was recorded at 311, 333, and 363 nm for cinnamoyl-CoA, 4-coumaroyl-CoA, and caffeoyl-CoA, respectively.

(D) Phylogenetic tree of SbCLL-7 and other members of the CLL-7 family from *Sesamum indicum*, *Salvia splendens*, and *Salvia miltiorrhiza* as well as SbCLL-1 and SbCLL-5, which are both 4-coumarate CoA ligases (Zhao et al., 2016b).

(E) Alignment of the active domains, which bind the adenylate acceptor, in CLL-7s from *Sesamum indicum* (Si), *Salvia miltiorrhiza* (Sm), *Salvia splendens* (Ss), *Scutellaria baicalensis* (Sb), and *Arabidopsis thaliana* (At). The C393 residue in SbCLL-7 (corresponding to C403 in At4CL2) is shown in bold.

caffeic acid. As shown in Figure 4A–4C, activity was observed for SbCLL-7 with cinnamic acid as the substrate, but this enzyme could not ligate CoA to 4-coumaric acid nor caffeic acid, in line with our previous observations (Zhao et al., 2016b). The syntenic CLL-7s from *S. miltiorrhiza*, *S. splendens*, and *S. indicum* had no activity with any of the aromatic substrates assayed, indicating that the specific activity of SbCLL-7 is not shared by the CoA ligases encoded by syntenic genes in species closely related to *S. baicalensis*.

SbCLL-7, SmCLL-7, SiCLL-7, and SsCLL-7 proteins were modeled using the Phyre2 server with PDB: 5BST (Nt4CL2 in thioester-forming conformation) as a template. Cinnamoyl-adenylate was docked into the active site using Autodock Vina. The CLL-7 protein folds largely overlapped, except in a loop region between Leu149 and Asp181 (Supplemental Figure 6A) (Li and Nair, 2015). The active site of SbCLL-7 is narrow and consists of hydrophobic residues, favoring a hydrophobic substrate such as cinnamoyl-adenylate (Supplemental Figures 6B and 5B) (Zhao et al., 2016b). The substrate is coordinated by hydrogen bonding with Thr343 and Asp424. The position of Phe246 suggests a π–π stacking interaction with the cinnamoyl head group (Supplemental Figure 6B). A cysteine residue (C393 corresponding to C403 in At4CL2) is present in SbCLL-7 but is substituted by more hydrophobic residues in the other CLL-7 proteins (Figure 4E) and might be responsible for SbCLL-7 binding an aromatic substrate (this residue lies in loop A6 and is also a cysteine residue in many 4-CoA ligases, Figure 4E). However, it was difficult to make any firm mechanistic conclusion from the structure model, because the C393 residue is fairly distant from the active site (Supplemental Figure 6C). The changes that led to SbCLL-7 being able to use cinnamate as a substrate must have involved point mutations of the *CLL-7* gene, <32.7 Mya (Zhao et al., 2016b).

Tandem Replication Contributed to Evolution of the RSF Pathway

Our previous work showed that *S. baicalensis* encodes two CHS isoforms, namely SbCHS-1 and SbCHS-2. The *SbCHS-1* gene is highly expressed in aerial parts, especially in flowers, and encodes an enzyme involved in classic flavone and anthocyanin biosynthesis (Zhao et al., 2016b). *SbCHS-2* transcripts are abundant in roots and encode an enzyme that produces

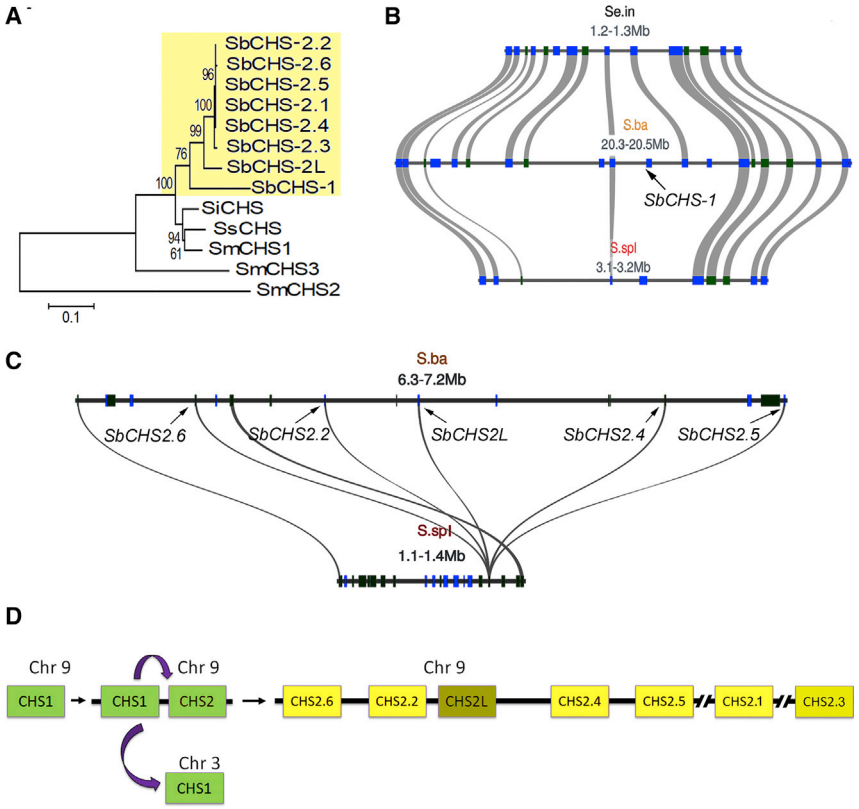

Figure 5. Evolution of *S. baicalensis CHS* Genes.

(A) Phylogenetic analysis of CHS proteins. Maximum-likelihood method was used to construct this tree with 1000 replicate bootstrap support. The tree was rooted with SmCHS2. SiCHS, XP_011091402.1; SsCHS, Saspl_016604-T1; SmCHS1, AWX67435.1; SmCHS2, AWX67436.1; SmCHS3, AWX67437.1.

(B) Syntenic analysis of SbCHS-1 in the related species *Sesamum indicum* (Se.in) and *Salvia splendens* (S.spl).

(C) Syntenic analysis of *SbCHS-2L* and *SbCHS-2* genes between *S. baicalensis* (S.ba) and *Salvia splendens* (S.spl).

(D) Diagram of proposed gene replication events that formed *CHS-2* genes in *S. baicalensis*. The classic *CHS* gene was originally located on chromosome 9 in the common ancestor of *Salvia* and *Scutellaria*. In the lineage giving rise to the genus *Scutellaria*, this gene was duplicated with one copy moving to pseudochromosome 3 (but maintaining its catalytic functionality) while the other copy underwent several duplications following neofunctionalization.

pinocembrin chalcone and is responsible for RSF biosynthesis (Zhao et al., 2016b).

In *S. baicalensis*, *CHS-1* is located on pseudochromosome 3 (Sb03g18491), although we did not find homologs of *SbCHS-1* in the isogenic regions of *S. indicum* and *S. splendens* (Figure 5A and 5B). In *S. miltiorrhiza*, the genome quality was not good enough for comparative analysis. In *S. baicalensis*, we found six loci encoding SbCHS-2 with percent identity ranging from 98.38% to 100% at the nucleotide level (Supplemental Table 15). We named these genes *SbCHS-2.1* to *SbCHS-2.6*. The coding sequences of *SbCHS-2.1*, *SbCHS-2.4*, and *SbCHS-2.5* were completely identical to the reported *SbCHS-2* sequence (Zhao et al., 2016b). Although *SbCHS-2.2* had two nucleotide differences to *SbCHS-2*, this did not lead to any amino acid differences in the predicted protein product compared with SbCHS-2. *SbCHS-2.6* and *SbCHS-2.3* differ from *SbCHS-2* by 1 and 19 nucleotides, respectively, and encode enzymes with just one and three amino acid differences, respectively, to the SbCHS-2 protein, suggesting functional specificity identical to that of SbCHS-2. The DNA sequence of Sb09g03140 shares 93% identity with *SbCHS-2* but only 85% identity with *SbCHS-1* at the nucleotide level, so this gene was named *SbCHS-2L*.

Genes encoding SbCHS-2.1 to SbCHS-2.6 and SbCHS-2L all locate on pseudochromosome 9, and interestingly, *SbCHS-2.2*, *SbCHS-2.3*, *SbCHS-2.4*, *SbCHS-2.5*, *SbCHS-2.6*, and *SbCHS-2L* are located close together, probably as a result of recent tandem replications, since they all share high similarity (Figure 5C and Supplemental Table 15). A corresponding region with *CHS* genes was not present in *S. indicum* but there was one *CHS* gene in the syntenic region of *S. splendens* encoding a protein that aligned well with the products of the single *CHS* (*CHS-1*) genes in the other species (Figure 5A). This suggested that in the ancestor of *S. baicalensis* and *Salvia* spp., a single *CHS* gene was located in the region equivalent to pseudochromosome 9. In *S. baicalensis* this *CHS* (ancestor of *CHS-1* and *CHS-2*) likely duplicated and one copy moved to pseudochromosome 3 (*CHS-1*) while the other copy mutated (*CHS-2L*) and further duplicated (*CHS-2.3*); after the divergence of *Scutellaria* from *Salvia*, following further sequence divergence and neofunctionalization, *CHS-2.3* was amplified by tandem duplications to produce the other *CHS-2s*. Using the *Salvia*/*Scutellaria* divergence time estimated by comparison of the whole-genome sequences (~32.7 Mya; Figure 5B) to calibrate the tree, a phylogeny of the *CHS* genes in *S. baicalensis* was visualized with Densitree (Supplemental Figure 7). The first divergence of *CHS-2* genes was estimated at around 19 Mya (*SbCHS-2L*), and the second around 12 Mya (*SbCHS-2.3*). The tandem replication (*CHS-2.1*, *CHS-2.2*, *CHS-2.4*, *CHS-2.5*, and *CHS-2.6*) appeared to be very recent, occurring probably within the last 1 Mya.

In *Scutellaria*, CHI activity is shared by both aerial and RSF pathways (Zhao et al., 2016b). In *S. baicalensis* we found that CHI was encoded by a single locus located on pseudochromosome 3 with the ID Sb03g24280.

S. baicalensis produces two isoforms of flavone synthase II (FNSII): FNSII-1 converts narigenin to apigenin in aerial parts of the plant, while FNSII-2 is specific for chrysin biosynthesis in roots (Zhao et al., 2016a). There are two genes encoding FNSII-1 proteins in the *S. baicalensis* genome, with ID numbers Sb06g05860 (*FNSII-1.1*) and Sb03g20730 (*FNSII-1.2*) located on

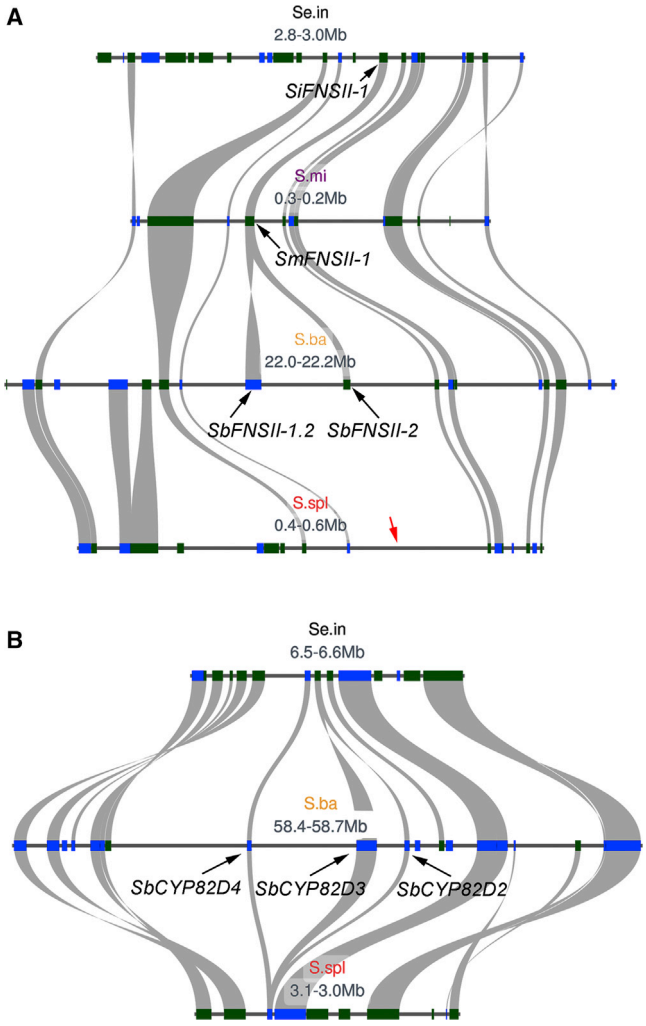

Figure 6. Tandem Duplications of *SbFNSII* and *SbCYP82D* Genes.

(A) Syntenic analysis of FNSII genes in *Sesamum indicum* (Se.in), *Salvia miltiorrhiza* (S.mi), *Salvia splendens* (S.spl), and *Scutellaria baicalensis* (S.ba).

(B) Syntenic analysis of CYP82D genes in *Sesamum indicum* (Se.in), *Salvia splendens* (S.spl), and *Scutellaria baicalensis* (S.ba).

pseudochromosome 6 and pseudochromosome 3, respectively. Both loci have open reading frames (ORFs) of 1509 bp and the two ORFs differ by just two nucleotides and one amino acid, suggesting they are the result of a relatively recent interchromosomal/segmental duplication (Figure 6A and Supplemental Figure 8). This was supported by analysis of the genes flanking *FNSII-1.1* and *FNSII-1.2* in *S. baicalensis* where four genes upstream of *FNSII-1.1* (Sb06g05990 to Sb06g05880) on pseudochromosome 6 are identical to the four genes upstream of *FNSII-1.2* (Sb03g20690 to Sb03g20720) on pseudochromosome 3. Downstream of *FNSII-1.1* on pseudochromosome 6 is a probable deletion of 13 genes before Sb06g05850, which is a duplicate of Sb03g20880 on pseudochromosome 3 (Supplemental Figure 9A). In *S. indicum* there is a single gene corresponding to *FNSII-1* in the region syntenic to *S. baicalensis* pseudochromosome 3, indicating that that the segmental duplication of part of pseudochromosome 3 to pseudochromosome 6 followed the divergence of the Lamiaceae from other families of the Lamiales (<46.2 Mya) (Figure 6A and Supplemental Figure 8).

In *S. splendens* there are two loci homologous to *SbFNSII-1* in the regions syntenic to *S. baicalensis* pseudochromosome 3 and pseudochromosome 6, confirming that the interchromosomal duplication occurred before the divergence of the family Lamiaceae (32.7 Mya) (Supplemental Figure 9B). However, the homolog of *FNSII-1.2* in *S. splendens* is a pseudogene that lacks an initiation codon, and consequently had not been annotated in the *S. splendens* genome sequence. *SbFNSII-2* (Sb03g20740) is specific to *S. baicalensis*. *SbFNSII-2* lies adjacent to *SbFNSII-1.2* on pseudochromosome 3 in *S. baicalensis*, in a tail-to-tail inverted orientation (Figure 6A). This observation further supports the idea that *SbFNSII-2* was produced by tandem duplication of *SbFNSII-1.2*, possibly as a result of non-homologous end-joining following double-stranded break-repair (Coen et al., 1986; Ballif et al., 2003), and neofunctionalization following divergence of the family Lamiaceae to form the genus *Scutellaria* (<32.7 Mya) (Zhao et al., 2016b).

In *S. baicalensis* SbCYP82D1 (encoded by Sb05g10371) carries out 6-hydroxylation of apigenin in the aerial parts and chrysin in the roots to produce scutellarein and baicalein, respectively. SbCYP82D2 is a flavone 8-hydroxylase and accepts only chrysin to make norwogonin in roots (Zhao et al., 2018). Although *SbCYP82D1* is located on pseudochromosome 5 in *S. baicalensis*, no sequence encoding a CYP82D enzyme could be found in the isogenic regions of *S. indicum* or *S. splendens* (Supplemental Figure 10) suggesting that *CYP82D1* may have relocated to pseudochromosome 5, perhaps from duplication of an original *CYP82D* gene on pseudochromosome 1, because the *SbCYP82D2* gene (ID Sb01g39830), which encodes flavone 8-hydroxylase, is located on pseudochromosome 1 in *S. baicalensis*. We found two other *CYP82D* genes in tandem upstream of *SbCYP82D2*, which we named *SbCYP82D3* (Sb01g39820) and *SbCYP82D4* (Sb01g39810). These genes shared nucleotide identities of 76% and 74%, respectively, with *SbCYP82D2* (Figure 6B). There are two genes encoding CYP82Ds in the syntenic region of *S. indicum* and one gene in *S. miltiorrhiza*.

Identification of Genes Encoding *O*-Methyltransferases with Potential Roles in Wogonin Biosynthesis

To investigate the genes responsible for the biosynthesis of wogonin, we screened the gene models annotated as OMTs from the *S. baicalensis* genome sequence, then performed BLAST searches against the database using sweet basil 8-OMT sequences as baits (Berim and Gang, 2013). This returned six genes encoding type II OMTs (Supplemental Table 16), which were closely related to the *PFOMT* (phenylpropanoid and flavonoid OMT) gene family, so we named them *SbPFOMT*s. We designed primer pairs for all the genes and successfully isolated five ORFs from *S. baicalensis* cDNA. No transcript for *SbPFOMT6* was detected in any of the tissues for which we obtained RNA-sequencing data. Information on the primers and the *OMT* genes is provided in Supplemental Table 17.

We expressed five *PFOMT* ORFs in yeast, and fed the cells with norwogonin. The strains were incubated overnight and

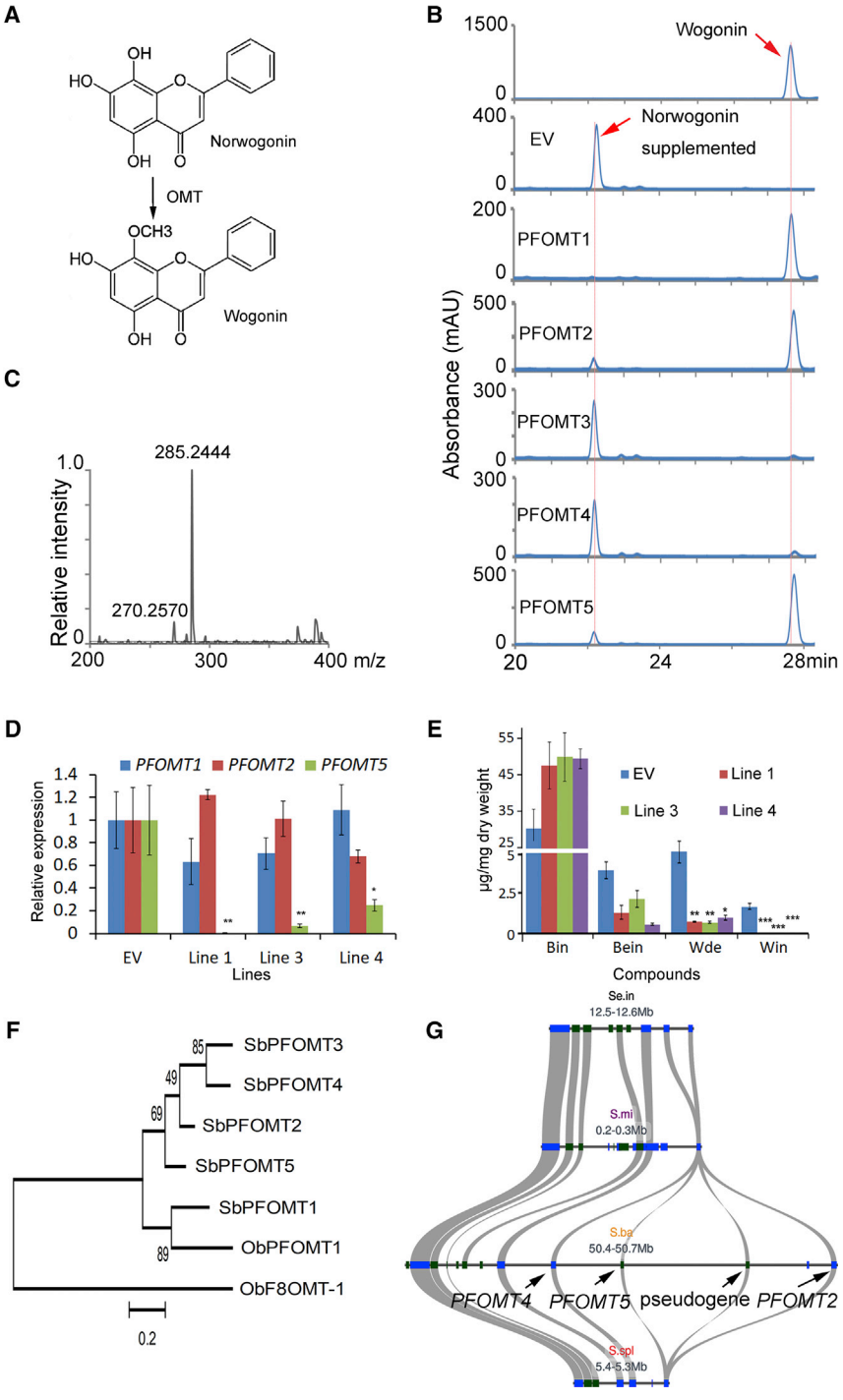

Figure 7. Screening of Candidate *PFOMT* Genes for Wogonin Biosynthesis.

(A) The proposed reaction carried out by the OMT that converts norwogonin to wogonin.

(B) HPLC analysis of yeast samples incubated with norwogonin as substrate *in vivo*. Top, wogonin standard; EV, yeast carrying the empty vector; PFOMT1–PFOMT5, yeast expressing the corresponding proteins, where a new peak with the same retention time as wogonin was found.

(C) MS II and fragmentation patterns of the new compound produced by PFOMTs expressed in yeast. The fragmentation patterns were identical to those of wogonin.

(D) Silencing of *SbPFOMT5* in RNAi hairy roots was measured by monitoring relative transcript levels by qRT–PCR.

(E) Measurements of RSFs from the *SbPFOMT5* RNAi lines. Bin, baicalin; Wde, wogonoside; Bein, baicalein; Win, wogonin. Standard errors were calculated from three biological replicates. $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ (Student's *t*-test).

(F) Phylogenetic tree of SbPFOMT proteins. Maximum-likelihood method was used to construct this tree with 1000 replicate bootstrap support. The tree was rooted with two *Oscimum basilicum* 8-OMTs: proteins encoded by *ObPFOMT-1* (KC354401) and *ObF8OMT-1* (KC354402).

(G) Tandem replication of *SbPFOMTs* was found on pseudochromosome 1 of *S. baicalensis*.

then extracted for liquid chromatography–mass spectrometry (LC–MS) analysis. All five of the yeast strains converted norwogonin into wogonin. PFOMT1, PFOMT2, and PFOMT5 could convert most of the norwogonin substrate into wogonin, whereas PFOMT3 and PFOMT4 produced very little wogonin (Figure 7A–7C). To assay the enzymes *in vitro*, we expressed the PFOMTs as 6×His-tag fusions in *E. coli* Rosetta, and proteins were extracted and purified to apparent homogeneity (Supplemental Figure 11). In accordance with *in vivo* results, all could convert norwogonin to wogonin *in vitro* in the presence of *S*-adenosylmethionine, although for PFOMT3 and PFOMT4 only a tiny peak of wogonin could be detected and only after extended periods of incubation, with excess added protein extract. The initial rate kinetic parameters for PFOMT1, PFOMT2, and PFOMT5 acting on norwogonin were determined and showed that PFOMT1, PFOMT2, and PFOMT5 had significantly different activities on the substrate. PFOMT5 had the greatest catalytic efficiency against norwogonin, with an apparent K_M of 4.58 μM and a maximal velocity of 10.34 nkat mg^{-1} (Supplemental Table 18). Compared with PFOMT5, PFOMT1 and PFOMT2 had lower activity on norwogonin as a result of their higher K_M values (29.08 μM and 11.98 μM, respectively) for norwogonin and their lower maximal velocities (8.07 nkat mg^{-1} and 7.75 nkat mg^{-1}, respectively), suggesting that PFOMT5 is the most efficient catalyst in the biosynthesis of wogonin in the roots of *S. baicalensis*.

Expression Patterns of the Genes Encoding PFOMTs Responsible for Wogonin Biosynthesis

The expression patterns of the *PFOMTs* were evaluated using the FPKM values (fragments per kilobase of transcript per million mapped reads) obtained from RNA-sequencing data. The transcript levels of *SbPFOMT1* were high specifically in stems of *S. baicalensis*, and the other *SbPFOMTs* were highly expressed

in roots or JA-treated roots, coinciding with the accumulation patterns of RSFs (Supplemental Figure 12).

Silencing of *PFOMTs* Identified *SbPFOMT5* as a Major Gene Contributing to Wogonin Synthesis in *S. baicalensis* Roots

Hairy root-mediated RNAi technology was employed to evaluate the contribution of the SbPFOMTs to wogonin biosynthesis. We found no differences in the four major flavones between the *PFOMT1* RNAi lines and an empty vector control line, even though in lines 1 and 4 the transcript levels of *PFOMT1* were reduced to 13% and 10%, respectively, compared with the control, as shown in Supplemental Figures 13 and 14. We screened three lines with different degrees of silencing of *PFOMT2* transcripts. qRT–PCR analysis showed two lines had significant downregulation of *PFOMT2*, with transcript levels 24% and 13% of those of the empty vector control (Supplemental Figures 12 and 13). However, wogonoside and wogonin levels were reduced only slightly in these lines (Supplemental Figures 13 and 14). We obtained three *PFOMT5* RNAi lines with dramatic reductions in PFOMT5 expression levels; *PFOMT5* transcript levels in line1, line 3, and line 4 were 0.6%, 7%, and 25%, respectively, of those in the control (Figure 7D). We found that the wogonin peak was completely absent in all three *SbPFOMT5* RNAi lines. There were also significant reductions in wogonoside, which was reduced from 5.16 mg/g dry weight in control roots to 0.73, 0.72, and 1.02 mg/g in line 1, line 3, and line 4, respectively (Figure 7E and Supplemental Figure 14). Consequently, both enzyme assays and RNAi demonstrated that *PFOMT5* encodes the enzyme with the highest catalytic efficiency for wogonin and wogonoside in the roots of *S. baicalensis*, and is likely the major enzyme catalyzing the synthesis of wogonin from norwogonin *in vivo*.

SbPFOMT2 (Sb01g34330), *SbPFOMT4* (Sb01g34290), *SbPFOMT5* (Sb01g34300), and *SbPFOMT6* (Sb01g34310) are clearly derived from tandem duplication (Figure 7F and 7G). There was just one PFOMT gene lying in the comparable DNA region in *S. miltiorrhiza* and S. *splendens*, while two copies were located in the syntenic region in *S. indicum* (Supplemental Figure 15A). In contrast, *SbPFOMT3* was unique to *S. baicalensis*, as no gene encoding a PFOMT was found in the syntenic region in any of the other genomes (Supplemental Figure 15B).

Based on our analyses, we suggest that four of the genes involved in RSF synthesis in *S. baicalensis* (*CHS-2*, *FNSII-2*, *F8H*, and *PFOMT5*) are the result of relatively recent tandem duplications, which were likely fundamental to the evolution of the pathway to synthesize RSFs.

DISCUSSION

Members of the family Lamiaceae are commonly known as mint plants. The family has a worldwide distribution, and includes between 6900 and 7200 species with the largest genera being *Salvia* (900 species), *Scutellaria* (360 species), and *Stachys* (300 species) (Raymond et al., 2004). This family of medicinal plants and herbs has a high degree of diversity in specialized metabolites and includes mint, rosemary, basil, thyme, marjoram, lavender, perilla, sage, and skullcaps (*Scutellaria*) (Wu, 1977). Terpenoids, phenolic acids, and flavonoids are mainly responsible for the bioactivities of these members of the family Lamiaceae (Wu et al., 2016; Zhao et al., 2016a). High-quality genome sequencing provides a key resource to study the molecular basis of the diversity in specialized metabolites in different members of the family Lamiaceae and how biosynthetic pathways evolved (Afendi et al., 2012). To date, only the genomes of *S. miltiorrhiza* and *S. splendens* from the family Lamiaceae have been reported (Xu et al., 2016; Dong et al., 2018). To these we can now add a high-quality 386.63 Mb (about 94%) reference genome sequence for *S. baicalensis*, the first for the genus *Scutellaria*. Since this is the first genome assembly at chromosome-level resolution in the family Lamiaceae, it should facilitate improved assembly of other genomes of members of the mint family, including *S. splendens* and *S. miltiorrhiza*. The genomic information reported offers a foundation for comparative genomic analysis between members of the family Lamiaceae and will facilitate elucidation of metabolic pathways, as well as molecular breeding, for this important medicinal plant.

Specific metabolites are important for the adaptation of plants to different environments, and these compounds are often lineage specific, strictly regulated, and highly evolvable (Weng et al., 2012). Evolution of specific metabolic pathways is driven by gene duplication and, subsequently, sub-/neofunctionalization (Panchy et al., 2016). RSFs, such as baicalein and wogonin, appear to be specific to the genus *Scutellaria* (Figure 1) (Liu et al., 2002; Xiao et al., 2003; Zhang et al., 2005; Makino et al., 2008; Lin and Shieh, 2015). Based on our comparative genomic analyses, we can now offer a description of how this specific metabolic pathway might have evolved.

The first enzyme committed to 4′-deoxyflavone biosynthesis is SbCLL-7, which ligates CoA specifically to cinnamate. The enzyme evolved from a CoA ligase involved in fatty acid/jasmonate biosynthesis (Schneider et al., 2005), and interestingly, none of the homeologous CLL-7 enzymes from *Sesamum* or *Salvia* had any activity on cinnamate, suggesting that the ability to synthesize high levels of 4′-deoxyflavones began with mutation of the ancestor of *SbCLL-7* such that the encoded enzyme could accept cinnamate as a substrate (<32.7 Mya). Members of this family of CoA ligases have an alanine residue in the binding pocket that binds the adenylate acceptor rather than the glutamine residue found in classic 4-CoA ligases, and the alanine allows binding of adenylate/CoA acceptors that are more hydrophobic than coumarate, such as cinnamate (Weng et al., 2012; Zhao et al., 2016b). However, the closely related, syntenous genes encoding CLL-7-like proteins in *Sesamum* and *Salvia* (which also have an alanine at the equivalent position in their active sites) are unable to accept cinnamate (Figure 4A–4C), indicating that other changes were necessary to secure the specificity of the encoded CoA ligase for cinnamate (Figure 4E and Supplemental Figure 5B). Notably, gene duplication was not involved in the evolution of the gene encoding this "signature" enzyme (Nützmann and Osbourn, 2014) of 4′-deoxyflavone biosynthesis in *Scutellaria*, providing an important example of evolution of a new metabolic pathway by increased enzyme promiscuity or adoption of new functionality, additional to the gene duplication and neo-/subfunctionalization

mechanism that has been elegantly described in recent genome papers on other species in the order Lamiales (Xu et al., 2016; Mint Evolutionary Genomics Consortium, 2018; Zhao et al., 2019).

The second gene involved in the synthesis of 4′-deoxyflavones is a gene encoding an isoform of CHS specific for cinnamoyl-CoA in combination with malonyl-CoA. The gene encoding the CHS active in the classic flavonoid pathway was likely originally located on pseudochromosome 9 in the common ancestor of *Salvia* and *Scutellaria*. In the lineage giving rise to the genus *Scutellaria*, this gene was likely duplicated with one copy moving to pseudochromosome 3 (but maintaining its catalytic functionality) while the other copy remained on pseudochromosome 9 and underwent several replications and subsequent neofunctionalization. Two duplications (19 Mya and 12 Mya) were followed by very recent multiplication of *SbCHS-2* to produce five gene copies encoding identical or near-identical proteins. This recent replication may represent an example of gene amplification to increase gene and protein dosage, perhaps to support greater flux along the 4′-deoxyflavone biosynthetic pathway (Figure 5D). Its occurrence within the last million years might reflect changes involving human selection for plants with higher levels of 4′-deoxyflavones in their roots for use in TCM.

The third enzyme in 4′-deoxyflavone biosynthesis is CHI, and this activity is encoded by a single gene in *S. baicalensis*, which is active in both the 4′-hydroxy (classic) and 4′-deoxyflavone biosynthetic pathways (Zhao et al., 2016b). This single-copy gene resides in equivalent positions in *S. indicum*, *S. splendens* and *S. baicalensis*, and has not undergone any significant changes associated with the evolution of the 4′-deoxyflavone pathway in *S. baicalensis*.

In *S. baicalensis* there are two *FNSII-1* loci and one *FNSII-2* locus. The genome of *S. splendens* also has two loci encoding FNSII-1, *FNSII-1.1* and *FNSII-1.2* (the *S. miltiorrhiza* genome sequence is of low quality in the area of *SmFNSII-1.1* and this gene could not be confirmed). The duplicated *SbFNSII-1.1*/*SbFNSII-1.2* genes have arisen as part of a segmental duplication between pseudochromosomes 6 and 3, which occurred after the divergence of the Lamiaceae (42.7 Mya). *S. splendens* has *FNSII-1.2* (although it is no longer functional) but no *FNSII-2*, suggesting *FNSII-2* was produced after divergence of *Salvia* from *Scutellaria* (<32.7 Mya) through tandem duplication of *FNSII-1.2* (Figure 5A; Supplemental Figures 8 and 9).

The gene encoding F6H (SbCYP82D1) is located on pseudochromosome 5 in *S. baicalensis* without any related genes around it. It likely moved to this position following the multiplication of *CYP82D* genes on pseudochromosome 1 (Supplemental Figure 10). Catalytic evidence suggests that F8H hydroxylase activity (encoded by *SbCYP82D2*) might have evolved from an ancestral gene encoding F6H activity (Zhao et al., 2018). The *SbCYP82D1* gene is surrounded by fragments of LTR retrotransposons, suggesting it acquired its new position either by unequal crossing over between retrotransposons or by retrotransposition. *F8H* (*SbCYP82D2*) is located in a tandem repeat (Figure 6B). There are two and one *CYP82D* genes in the corresponding areas of *S. indicum* and *S. splendens*, respectively, supporting the view that relatively frequent duplications of these genes results in new decorating activities for flavonoids in the mint family (Figure 6B).

Overall, we found that four of the RSF genes (*CHS-2*, *FNSII-2*, *F8H*, and *PFOMT5*) are located in duplicated regions, which may have arisen by whole-genome duplication, tandem duplication through unequal crossing over, transposon-mediated gene duplication, segmental duplication, and retroduplication (Panchy et al., 2016). Tandem duplication and transposon-mediated gene duplication were likely involved in *CHS-2* recruitment, segmental duplication and then tandem or transposon-mediated duplication were probably involved in *FNSII-2* recruitment, tandem duplication and retroduplication were likely involved in *F8H* and *F6H* recruitment, and tandem duplications involving unequal crossing over were likely involved in recruitment of *PFOMT5* to wogonin biosynthesis.

We found no strong evidence for clustering of genes encoding different enzymes of this newly derived pathway. The gene encoding the key signature enzyme, SbCLL-7, lies on pseudochromosome 9 as do the genes encoding the enzyme involved in the next step, CHS-2. Perhaps this is evidence of gene clustering that facilitated co-evolution of the earliest steps of 4′-deoxyflavone biosynthesis in polymorphic populations, although the linkage is not tight (circa 12 000 genes separating the two loci) and does not serve as evidence of gene clustering under current criteria (Zhao et al., 2019).

The 4′-deoxyflavone pathway evolved relatively recently (<32.7 Mya) specifically in the genus *Scutellaria* by genome changes involving mutation (*SbCLL-7*), gene duplication and neofunctionalization (*SbCHS-2*, *SbFNSII-2*, *SbF8H*, and *SbPFOMT*), gene amplification (*SbCHS-2*), segmental duplication (*SbFNSII-1*), tandem duplication and neofunctionalization (*CHS-2*, *FNSII-2*), and tandem duplication and subfunctionalization (*SbPFOMT5*). Interestingly in the two cases of neofunctionalization of genes encoding enzymes of flavone biosynthesis (*CHS1/2* and *F6H/F8H*), the gene encoding the isoform active in the classic flavone pathway has a different chromosomal localization to the gene(s) encoding the isoforms active in the specialized 4′-deoxyflavone pathway, despite there being phylogenetic evidence that these isoforms were derived from duplications. Perhaps this separation of chromosomal locations allows more rapid functional divergence. This is not the case for *FNSII-1.2* and *FNSII-2* in *S. baicalensis*, although there is a second copy of the *FNSII-1* gene (*FNSII-1.1*) lying on a separate chromosome as a result of segmental duplication. Analysis of the new reference genome of *S. baicalensis* and comparison with high-quality genome sequences of two closely related species has allowed us to propose how this new metabolic pathway evolved and has revealed that a range of different evolutionary mechanisms have resulted in the emergence of a specialized metabolic pathway within a single genus. Reciprocal comparisons with high-quality genomes of other members of the mint family should similarly illuminate the evolution of specialized biosynthetic pathways for terpenoids such as the diterpenoid tanshinone (Xu et al., 2016), phenolic acids, and flavonoids in this metabolically diverse family of plants.

METHODS

Genome Sequencing

Genomic DNA was extracted from leaves of a single *S. baicalensis* plant maintained in Shanghai Chenshan Botanical Garden, using a modified CTAB method (Tel-Zur et al., 1999). Quality control was done using a Sage Science Pippin pulse electrophoresis system. Genomic DNA with a length of around 150 kb was sheared using a Megaruptor DNA system and the resulting fragments of 30–50 kb were collected for the following steps. A SMRTbell DNA library was constructed with Sequel 2.0 reagent according to the manufacturer's instructions (Pacific Bioscience, www.pacb.com). SMRTbell sequencing primers together with P4 DNA polymerase were used for sequencing reactions on SMART Cells. The genomic DNA was sequenced on the PacBio Sequel system. This work produced 48.02 Gb of single-molecule data representing about 117.66× coverage of the genome (Supplemental Table 1).

For short-insert library construction, DNA was extracted from the same plant using a DNA secure plant kit (Tiangen, http://www.tiangen.com/) according to the manufacturer's instructions. The DNA was sheared and fragments with sizes of 200–300 bp were retrieved from agarose gels. The fragments were ligated to adaptors and were selected for RNA amplification for templates. The library was then sequenced on HiSeq X Ten. The Illumina sequencing produced 67.96 Gb of short-reads data, amounting to about 166.51× coverage of the genome.

For 10x Genomics sequencing, we extracted DNA samples using the modified CTAB method (Tel-Zur et al., 1999). The GemCode Instrument (10x Genomics) was employed for DNA indexing and barcoding. GEM reactions were carried out using about 1-ng of 50-kb single DNA molecules, and 16-bp barcodes were ligated to the molecules in droplets. The intermediate DNA was extracted from the droplets and sheared to 500-bp fragments (Zheng et al., 2016). The fragments were ligated to P7 adaptors, which were then sequenced on an Illumina HiSeq X Ten platform (Mostovoy et al., 2016). We obtained 86.72 Gb of data from 10x Genomics sequencing, which were then used for genome assembly.

DNA from young leaves of the same *S. baicalensis* plant was used as starting material for the Hi-C library. Formaldehyde was used for fixing chromatin. The leaf cells were lysed and DpnII endonuclease was used for digesting the fixed chromatin. The 5′ overhangs of the DNA were recovered with biotin-labeled nucleotides and the resulting blunt ends were ligated to each other using DNA ligase. Proteins were removed with protease to release the DNA molecules from the crosslinks. The purified DNA was sheared into 350-bp fragments and ligated to adaptors (Yaffe and Tanay, 2011). The fragments labeled with biotin were extracted using streptavidin beads and after PCR enrichment, the libraries were sequenced on Illumina HiSeq PE150.

Estimation of Genome Size

The genome size was measured by flow cytometry according to the protocol described by Doležel and Bartoš (2005). In brief, young leaves of *S. baicalensis* were chopped with a sharp razor blade for 60 s in nuclear isolation buffer (200 mM Tris, 4 mM $MgCl_2 \cdot 6H_2O$, 0.5% [v/v] Triton X-100 [pH 7.5]). Samples were incubated for 3 min and filtered through 50-μm CellTrics filters. Plant cell nuclei were stained by adding 2 ml of buffer containing propidium iodide and RNAse A in the dark for 2 min. The relative nuclear genome size was analyzed on a flow cytometer (BD FACSAria III). This analysis gave us an estimated genome size of 392 Mb compared with the tomato genome (Tomato Genome Consortium, 2012). We evaluated further the genome size by performing k-mer frequency analysis based on Illumina short reads using the K-mer Analysis Toolkit (http://www.earlham.ac.uk/kat-tools).

Genome Assembly

De novo assembly was carried out using PacBio reads. Error correction was conducted by mapping the seed reads in FALCON according to the manufacturer's instructions (https://github.com/PacificBiosciences/FALCON/wiki/Manual) with the following parameters: –max_diff 100; –max_cov 100; –min_cov 2; –min_len 5000. Contaminants were removed with PacBio's whitelisting pipeline. The resulting primary assembly was phased using FALCON-Unzip (default parameters). Heterozygosity of the contigs was analyzed with FALCON-Unzip, which were then phased according to the differences. The phased sequences were assembled into haplotigs for a diploid. The contigs were polished with Quiver (http://pbsmrtpipe.readthedocs.io/en/master/getting_started.html) with the parameters: pbsmrtpipe.options.chunk_mode: Truepbsmrtpipe.options.max_nchunks:50 to produce primary contigs, which were further corrected with reference to the Illumina reads with Pilon (https://github.com/broadinstitute/pilon/wiki).

RNA Sequencing and Analysis

Six tissues were harvested from 3-month-old *S. baicalensis* plants, namely flower buds, flowers, leaves, roots, JA-treated roots (100 μM MeJA treated for 24 h), and stems. Three biological replicates for each tissue were collected. Total RNA was extracted from these tissues using the RNAprep pure plant kit (Tiangen). After removing DNA, mRNA was isolated using oligo(dT) beads. The mRNA was harvested and broken into short fragments, which were then used as templates for cDNA synthesis. After end repair, single-nucleotide A (adenine) addition and adapter ligation to the cDNA were undertaken. Fragments of 200–300 bp were separated for PCR amplification. An Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System were used for quality control of the libraries, which were then sequenced on Illumina HiSeq 2000. Raw reads produced by sequencing were stored in Fastq format. Raw reads with adaptors and unknown nucleotides and low-quality reads with more than 20% low-quality bases (base quality $\leq$10) were removed to leave clean reads. Trinity (Haas et al., 2013) was employed for *de novo* assembly based on the clean reads to produce Unigenes.

To annotate the Unigenes, we carried out blast (blastX and blastn) searches against the NR, SwissProt, KEGG, COG, and NT databases (E value <0.00001). Annotations of the proteins with the highest similarity to each Unigene were retrieved. GO annotations based on similarity were then assigned to each Unigene using Blast2GO (Conesa et al., 2005). Expression of Unigenes was calculated using the FPKM method (Mortazavi et al., 2008). The formula was: $FPKM = 10^6 C/NL/10^3$, where C is the number of fragments that aligned specifically to a certain Unigene, N is the total number of fragments that aligned to all Unigenes, and L is the number of bases in the CDS of the Unigene.

Evaluation of Genome Quality

To evaluate the coverage of the assembly, we mapped all the paired-end Illumina short reads to the assembly using BWA (http://bio-bwa.sourceforge.net/) (Li and Durbin, 2009). Gene completeness was evaluated using the ESTs generated from RNA sequencing. The ESTs were mapped to the assembly using BLAT (http://genome.ucsc.edu/goldenpath/help/blatSpec.html). For CEGMA (http://korflab.ucdavis.edu/dataseda/cegma/) evaluation, we built a set of highly reliable conserved protein families that occur in a range of model eukaryotes (Parra et al., 2007). We then mapped the 248 core eukaryotic genes to the genome. The genome was also assessed using the BUSCO (http://busco.ezlab.org/) gene set (Simão et al., 2015), which includes 956 single-copy orthologous genes.

Genome Annotation

Repeat elements were annotated using a combined strategy. Alignment searches were undertaken against the RepBase database (http://www.girinst.org/repbase), then Repeatproteinmask searches (http://www.repeatmasker.org/) were used for prediction of homologs

(Jurka et al., 2005). For *de novo* annotation of repeat elements, LTR_FINDER (http://tlife.fudan.edu.cn/ltr_finder/), Piler (http://www.drive5.com/piler/), RepeatScout (http://www.repeatmasker.org/), and RepeatModeler (http://www.repeatmasker.org/RepeatModeler/html/) were used to construct a *de novo* library, then annotation was carried out with Repeatmasker (Price et al., 2005). Non-coding RNA was annotated using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) (for tRNA) or INFERNAL (http://infernal.janelia.org/) (for miRNA and snRNA). Since rRNA sequences are highly conserved among plants, rRNAs from *S. baicalensis* were identified by blast searches.

Gene structure screening was carried out through a combination of homology, *de novo*, and EST-based predictions. A gene set including protein coding sequences from *A. thaliana*, *Oryza brachyantha*, *P. trichocarpa*, *S. miltiorrhiza*, *Sesamum indicum*, *S. lycopersicum*, and *Utricularia gibba* was mapped to the assembly of *S. baicalensis* using blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (E value $\leq 1e^{-5}$), and the gene structure of each hit was predicted through Genewise (http://www.ebi.ac.uk/~birney/wise2/) (Birney et al., 2004). *De novo* gene structure identification was performed using Augustus (http://bioinf.uni-greifswald.de/augustus/) and GlimmerHMM (http://ccb.jhu.edu/software/glimmerhmm/). Gene structure was also determined by mapping ESTs to the assembly through BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). All the resulting genes were corrected by reference to the transcriptome data and integrated into a non-redundant gene set using EVidenceModeler (http://evidencemodeler.sourceforge.net/) (Haas et al., 2008). The genes were further corrected with PASA (Program to Assemble Spliced Alignment; http://pasa.sourceforge.net/) to predict untranslated regions and alternative splicing (Haas et al., 2008), which generated 28 930 gene models.

Gene function was annotated by performing BLASTP (E value $\leq 1e^{-5}$) searches against the protein databases SwissProt (http://www.uniprot.org/), TrEMBL (http://www.uniprot.org/), and KEGG (http://www.genome.jp/kegg/). InterPro (https://www.ebi.ac.uk/interpro/) and Pfam (http://pfam.xfam.org/) were used for screening the functional domains of the proteins. GO terms based on the search were then assigned to each gene.

Gene Isolation

The ORFs of the OMT genes (*SbPFOMT1*, *SbPFOMT2*, *SbPFOMT3*, *SbPFOMT4*, and *SbPFOMT5*) were isolated using the primers listed in Supplemental Table 17, designed based on the genome and RNA sequencing data. The fragments were inserted into plasmid pDONR207 using Gateway BP Clonase II Enzyme Mix (http://www.thermofisher.com). Full-length CDSs (coding sequences) of the OMTs were then cloned into pYesdest52 (for all the *PFOMT*s) and pDEST17 (for *CLL-7s* and *PFOMT*s) for yeast and *E. coli* expression, respectively, using Gateway LR Clonase II Enzyme Mix (http://www.thermofisher.com) according to the manufacturer's instructions.

Protein Expression and Enzyme Assays

Vectors were constructed using the yeast plasmid pYesdest52 together with the isolated *OMT* genes. These together with an empty vector control were transformed into yeast *Saccharomyces cerevisiae* BY4742 for expressing the proteins. After culture for 2 days at 28°C, the transformants were screened on synthetic dropin medium -Ura (SD-Ura) supplemented with 20 g l^{-1} glucose. Liquid cultures of the engineered strains were set up by picking a single colony and growing in 10 ml of SD-Ura liquid medium with 20 g l^{-1} glucose at 28°C overnight. The cells were harvested and resuspended in SD-Ura medium supplemented with 20 g l^{-1} galactose to induce expression of the target proteins. Meanwhile, norwogonin was added at 20 μM to the cultures. After 16 h of cultivation the cells were centrifuged, harvested, and extracted with 70% MeOH for LC–MS analysis.

The constructs for the four *CLL-7* genes and the five *PFOMTs* in pDest17 were used for transformation of *E. coli* strain Rosetta. Single colonies of each transformant were grown overnight in 5 ml of Luria–Bertani (LB) medium with 100 mg l^{-1} ampicillin at 37°C. The strains were inoculated into 200 ml of fresh LB medium in the presence of 0.5 mM isopropyl-1-thio-L-D-galactopyranoside and grown at 16°C for 16 h to induce expression of the target proteins. The cells were harvested by centrifugation and suspended in 5 ml of buffer A (50 mM sodium phosphate [pH 7.8] containing 300 mM NaCl, 2 mM β-mercaptoethanol, 20% glycerol, 10 mM imidazole). The cells were lysed by sonication and cell debris was removed by centrifugation (12 000 *g*, 10 min, 4°C). Nickel–nitrilotriacetic acid (Ni-NTA) agarose (1 ml) (Qiagen, Hilden, Germany) was added to the supernatant, and the lysate was incubated with shaking at 4°C for 1 h for binding the His6-tagged proteins. The Ni-NTA agarose was washed with 30 ml of buffer B (50 mM sodium phosphate [pH 7.8] containing 300 mM NaCl, 20 mM imidazole, 15% glycerol, 2 mM β-mercaptoethanol). Subsequently the agarose was packed into a column and the protein was eluted with buffer C (50 mM sodium phosphate [pH 7.8] containing 300 mM NaCl, 250 mM imidazole, 20% glycerol, 2 mM β-mercaptoethanol). Fractions (1 ml) were collected and imidazole was removed by ultrafiltration. Concentrations of the eluted proteins were measured using the Bradford method (Bradford, 1976). The proteins were further analyzed by SDS–PAGE and western blotting (Mahmood and Yang, 2012).

CLL-7 enzymes were assayed in a system of 0.3 μM CoA, 5 μM $MgCl_2$, 50 mM Tris–HCl (pH 7.8), 5 μM ATP, and protein in 100 μl. PFOMT enzymes were assayed in 0.1 mM Tris–HCl (pH 7), 0.1 mM $MgCl_2$, and 5 mM S-adenosyl methionine in 100 μl of reaction system. The reactions were incubated at 28°C for 10 min (PFOMT5) or 30 min (PFOMT1 and PFOMT2). Norwogonin was used as a substrate at concentrations ranging from 0.1 to 4 mM. The kinetic constants, K_M and V_{max}, were determined by linear regression of *v* against *v/s* (Eadie–Hofstee plot). Each plot contained at least seven points.

Hairy Root-Mediated RNAi

For RNAi of candidate OMTs, we amplified non-conserved regions of *SbPFOMT1*, *SbPFOMT2*, and *SbPFOMT5* using the primers listed in Supplemental Table 17. The fragments were then subcloned into pDONR207 using Gateway recombination. After confirmation by DNA sequencing, the fragments were recombined into vector pK7WGIGW2R. The resulting RNAi constructs were used for transformation of *Agrobocterium rhizogenes* strain A4 by electroporation. Positive transformants were screened on TY medium supplemented with 50 mg l^{-1} kanamycin and 100 mg l^{-1} spectromycin. The engineered strains were grown overnight at 28°C, harvested by centrifugation (4000 *g* at 4°C), and resuspended in MS liquid medium to make *A. rhizogenes* suspension solutions.

Leaf explants were collected from 6-month-old *S. baicalensis* plants. The explants were sterilized with 10% bleach for 10 min then and washed with sterile water three to five times. A sharp knife was dipped into an *A. rhizogenes* suspension and then used to scratch the leaf explants, which were then dried on sterile filter paper and co-cultured on B5 medium containing 50 μM acetosyringone in the dark at 25°C. After 3 days of co-cultivation, the explants were transferred to B5 medium supplemented with 500 mg l^{-1} cefotaxime (Sigma). Several weeks later, hairy roots could be found at the wound site over the leaf veins. Successful transformants were screened using a fluorescence microscope, as pK7WGIGW2R carries a DsRed gene. The DsRed-positive hairy roots were removed from explants and kept on plates as independent lines.

Liquid hairy root cultures were established by growing a root tip of each line in 10 ml of B5 liquid with 400 mg l^{-1} cefotaxime in flasks, and the cultures were maintained at 25°C with shaking (90 rpm). During the growth of the hairy roots, B5 medium was added gradually up to 50 ml. Hairy roots were collected after 50 days, ground into powder in liquid N_2, and freeze-dried. One gram of the samples was extracted by sonication in 1

ml of 70% methanol for 2 h. After filtering through 0.22-μm columns, the extracts were used for LC–MS analysis.

Metabolite Analyses

An Agilent 1260 Infinity II HPLC (high-performance liquid chromatography) system was used for metabolite analysis. Separation was carried out on a 100 × 2-mm 3μ Luna C18(2) column using 0.1% formic acid in water (A) versus 1:1 acetonitrile/MeOH + 0.1% formic acid (B) and run at 260 μl min^{-1} with the following gradient: 0–3 min, 20% B; 20 min, 50% B; 20–30 min, 50% B; 36 min, 30% B; 37 min, 20% B; and 37–43 min, 20% B. The column was maintained at 35°C and absorption was detected at 280 nm with a diode array detector (Agilent). Metabolites were measured by comparing the area of the individual peaks with standard curves obtained from standard compounds.

LC–MS/MS was carried out on an IT-ToF mass spectrometer attached to a Prominence/Nexera UHPLC system (Shimadzu). Separation was on a 100 × 2-mm 3μ Luna C18(2) column using the same gradient described previously. Flavonoids were detected by UV absorption, collecting spectra from 200–600 nm from which we extracted chromatograms at the desired wavelength, 280 nm. They were also detected by positive electrospray MS, collecting spectra from *m/z* 200–2000 (using automatic sensitivity control, 70% of the base peak chromatogram). The instrument also collected automatic (data-dependent) MS2 spectra of the most abundant precursor ions, at an isolation width of *m/z* 3.0, 50% collision energy, and 50% collision gas, with an ion accumulation time of 10 ms. Spray chamber conditions were 250°C curved desorbation line, 300°C heat-block, 1.5 l min^{-1} nebulizing gas, and drying gas "on." The instrument was calibrated using sodium trifluoroacetate before use.

Statistics

All experiments were repeated using at least three biological replicates. Data are presented as means ±standard error of the mean, unless stated otherwise. To compare group differences, paired or unpaired, we used two-tailed Student's *t*-tests. *P* values of less than 0.05 were recognized as significant.

ACCESSION NUMBERS

Reference genome data are deposited in GenBank under project number PRJNA484052, and transcriptome sequence reads are deposited in the Sequence Read Archive (SRA) under accession number SRA: SRP156996.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the National Key R&D Program of China (2018YFC1706200, 2018YFD1000701-4), the National Natural Science Foundation of China (31870282, 31700268 and 31788103), the Fund of Chinese Academy of Sciences (QYZDY-SSW-SMC026 and 153D31KYSB20160074), the Chenshan Special Fund for Shanghai Landscaping Administration Bureau Program (G182401, G172402, G182402, G192413, and G192414), and the CAS/JIC and Center of Excellence for Plant and Microbial Sciences (CEPAMS) joint foundation through support to Q.Z., X.Y.C., J.Y., and C.M. Q.Z. and J.Y. were also supported by the Youth Innovation Promotion Association, Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

Q.Z. and C.M. initiated the program, coordinated the project, and wrote the manuscript. L.J., M.-Y.C., Y.F., Y.W., J.L., and L.Y. prepared and analyzed the samples. Q.Z., W.Q., Z.X., R.Y., and H.S. designed the sequencing strategy and performed the sequencing. J.Y., Q.Z., M.Y., J.-K.W., T.P., M.V., B.S., X.-Y.C., Y.H., and C.M. performed the analyses of the genome sequence.

ACKNOWLEDGMENTS

We thank Hong Zhu of the Core Facility for Plant Cell Biology, Shanghai Center for Plant Stress Biology, Chinese Academy of Sciences for her help in the operation of the flow cytometer, and Allan Downie for helpful comments on the manuscript. No conflict of interest declared.

Received: February 2, 2019
Revised: March 20, 2019
Accepted: April 7, 2019
Published: April 14, 2019

REFERENCES

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L.K., et al. (2012). KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. **53**:e1.

***Arabidopsis* Genome Initiative.** (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Ballif, B.C., Yu, W., Shaw, C.A., Kashork, C.D., and Shaffer, L.G. (2003). Monosomy 1p36 breakpoint junctions suggest pre-meiotic breakage–fusion–bridge cycles are involved in generating terminal deletions. Hum. Mol. Genet. **12**:2153–2165.

Berim, A., and Gang, D.R. (2013). Characterization of two candidate flavone 8-O-methyltransferases suggests the existence of two potential routes to nevadensin in sweet basil. Phytochemistry **92**:33–41.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. Genome Res. **14**:988–995.

Bradford, M.M. (1976). A rapid method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. **72**:248–254.

Cheng, Z., Cai, M., Hao, D., Deng, R., and Yan, L. (2010). Karyotype analysis and meiotic observations of pollen mother cells in *Scutellaria baicalensis* Georgi. Chinese Wild Plant Resources **2**:34–37, 58.

Coen, E.S., Carpenter, R., and Martin, C. (1986). Transposable elements generate novel spatial patterns of gene expression in *Antirrhinum majus*. Cell **47**:285–296.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**:3674–3676.

Doležel, J., and Bartoš, J. (2005). Plant DNA flow cytometry and estimation of nuclear genome size. Ann. Bot. **95**:99–110.

Dong, A.X., Xin, H.B., Li, Z.J., Liu, H., Sun, Y.Q., Nie, S., Zhao, Z.N., Cui, R.F., Zhang, R.G., and Yun, Q.Z. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. Gigascience **7**. https://doi.org/10.1093/gigascience/giy068.

Gao, J., Morgan, W.A., Sanchez-Medina, A., and Corcoran, O. (2011). The ethanol extract of *Scutellaria baicalensis* and the active compounds induce cell cycle arrest and apoptosis including upregulation of p53 and Bax in human lung cancer cells. Toxicol. Appl. Pharmacol. **254**:221–228.

Haas, B.J., Salzberg, S.L., Wei, Z., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. **9**:R7.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., and Lieber, M. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. **8**:1494–1512.

Islam, M.N., Downey, F., and Ng, C.K.Y. (2011). Comparative analysis of bioactive phytochemicals from *Scutellaria baicalensis*, *Scutellaria lateriflora*, *Scutellaria racemosa*, *Scutellaria tomentosa* and *Scutellaria wrightii* by LC-DAD-MS. Metabolomics **7**:446–453.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.

Kovács, G., Kuzovkina, I.N., Szoke, É., and Kursinszki, L. (2004). HPLC determination of flavonoids in hairy-root cultures of *Scutellaria baicalensis* Georgi. Chromatographia **60**:S81–S85.

Li-Weber, M. (2009). New therapeutic aspects of flavones: the anticancer properties of *Scutellaria* and its main active constituents Wogonin, Baicalein and Baicalin. Cancer Treat. Rev. **35**:57–68.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics **25**:1754–1760.

Li, S. (2012). Compendium of Materia Medica (Bencao Gangmu) (Beijing: Huaxia Press), pp. 543–546, first published 1593, republished 2012.

Li, Z., and Nair, S.K. (2015). Structural basis for specificity and flexibility in a plant 4-Coumarate:CoA ligase. Structure **23**:2032–2042.

Lin, C.C., and Shieh, D.E. (2015). In vivo hepatoprotective effect of baicalein, baicalin and wogonin from *Scutellaria rivularis*. Phytother. Res. **10**:651–654.

Liu, M., Yang, L., Wan, Y., and Zuo, F. (2002). Determination of baicalin and wogonoside in seven species of radix Scutellariae by RP-HPLC. Chin. J. Pharm. Anal., 2002-02.

Liu, X., Liu, Y., Huang, P., Ma, Y., Qing, Z., Tang, Q., Cao, H., Cheng, P., Zheng, Y., Yuan, Z., et al. (2017). The genome of medicinal plant *Macleaya cordata* provides new insights into benzylisoquinoline alkaloids metabolism. Mol. Plant **10**:975–989.

Ma, J.X. (2013). Explanatory Notes to Shennong Bencao Jing, 3 (Beijing: People's Medical Publishing House), p. 140.

Mahmood, T., and Yang, P.C. (2012). Western blot: technique, theory, and trouble shooting. N. Am. J. Med. Sci. **4**:429–434.

Makino, T., Hishida, A., Goda, Y., and Mizukami, H. (2008). Comparison of the major flavonoid content of *S. baicalensis*, *S. lateriflora*, and their commercial products. Nat. Med. **62**:294–299.

Mint Evolutionary Genomics Consortium. (2018). Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. Mol. Plant **11**:1084–1096.

Mochida, K., Sakurai, T., Seki, H., Yoshida, T., Takahagi, K., Sawai, S., Uchiyama, H., Muranaka, T., and Saito, K. (2017). Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. Plant J. **89**:181–194.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **5**:621–628.

Mostovoy, Y., Levysakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., and Džakula, Ž. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. Nat. Methods **13**:587–590.

Nützmann, H.W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. Curr. Opin. Biotechnol. **26**:91–99.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. (2016). Evolution of gene duplication in plants. Plant Physiol. **171**:2294–2316.

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

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21** (*Suppl 1*):i351.

Qiao, X., Li, R., Song, W., Miao, W.J., Liu, J., Chen, H.B., Guo, D.A., and Ye, M. (2016). A targeted strategy to analyze untargeted mass spectral data: rapid chemical profiling of *Scutellaria baicalensis* using ultra-high performance liquid chromatography coupled with hybrid quadrupole orbitrap mass spectrometry and key ion filtering. J. Chromatogr. A **1441**:83–95.

Raymond, M., Harley, S.A., Budantsev, A.L., Cantino, P.D., Conn, B.J., Grayer, R.J., Harley, M.M., de Kok, R.P.J., Krestovskaja, T.V., Morales, R., et al. (2004). The Families and Genera of Vascular Plants, VII (Berlin, Heidelberg: Springer-Verlag).

Schneider, K., Kienow, L., Schmelzer, E., Colby, T., Bartsch, M., Miersch, O., Wasternack, C., Kombrink, E., and Stuible, H.P. (2005). A new type of peroxisomal acyl-coenzyme A synthetase from *Arabidopsis thaliana* has the catalytic capacity to activate biosynthetic precursors of jasmonic acid. J. Biol. Chem. **280**:13962–13972.

Shang, X.F., He, X.R., He, X.Y., Li, M.X., Zhang, R.X., Fan, P.C., Zhang, Q.L., and Jia, Z.P. (2010). The genus *Scutellaria* an ethnopharmacological and phytochemical review. J. Ethnopharmacol. **128**:279–313.

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.

Tel-Zur, N., Abbo, S., Myslabodski, D., and Mizrahi, Y. (1999). Modified CTAB procedure for DNA isolation from epiphytic cacti of the. Plant Mol. Biol. Rep. **17**:249–254.

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

Tu, B., Li, R.R., Liu, Z.J., Chen, Z.F., Ouyang, Y., and Hu, Y.J. (2016). Structure-activity relationship study between baicalein and wogonin by spectrometry, molecular docking and microcalorimetry. Food Chem. **208**:192–198.

Wen, J. (2007). Sho-saiko-to, a clinically documented herbal preperation for treating chronic liver disease. HerbalGram **73**:34–43.

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

Wu, Y.B., Ni, Z.Y., Shi, Q.W., Dong, M., Kiyota, H., Gu, Y.C., and Cong, B. (2016). Constituents from *Salvia* species and their biological activities. Chem. Rev. **112**:5967–6026.

Wu, Z.-Y.L. (1977). Flora of China (Beijing: Science Press), pp. 194–198.

Xiao, L.H., Wang, H.Y., Song, S.J., Zhang, G.P., Song, H.X., and Sui, X.U. (2003). Isolation and identification of the chemical constituents of roots of *Scutellaria amoena* C.H. Wright. J. Shenyang Pharm. Univ. **20**:181–183.

Xu, H., Song, J., Luo, H., Zhang, Y., Li, Q., Zhu, Y., Xu, J., Li, Y., Song, C., Wang, B., et al. (2016). Analysis of the genome sequence of the medicinal plant salvia miltiorrhiza. Mol. Plant **9**:949–952.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. **43**:1059.

Yang, M.D., Chiang, Y.M., Higashiyama, R., Asahina, K., Mann, D.A., Mann, J., Wang, C.C.C., and Tsukamoto, H. (2012). Rosmarinic acid and baicalin epigenetically derepress peroxisomal proliferator-activated receptor gamma in hepatic stellate cells for their antifibrotic effect. Hepatology **55**:1271–1281.

Zhang, C., Zhang, Y., Chen, J., and Liang, X. (2005). Purification and characterization of baicalin-β-d-glucuronidase hydrolyzing baicalin to baicalein from fresh roots of *Scutellaria viscidula* Bge. Process Bioch. **40**:1911–1915.

Zhang, G., Tian, Y., Zhang, J., Shu, L., Yang, S., Wang, W., Sheng, J., Dong, Y., and Chen, W. (2015). Hybrid de novo genome assembly of

the Chinese herbal plant danshen (*Salvia miltiorrhiza* Bunge). GigaScience **4**:62.

Zhang, H., Miao, H., Lei, W., Qu, L., Liu, H., Qiang, W., and Yue, M. (2013). Genome sequencing of the important oilseed crop *Sesamum indicum* L. Genome Biol. **14**:401.

Zhao, D., Hamilton, J.P., Bhat, W.W., Johnson, S.R., Godden, G.T., Kinser, T.J., Boachon, B., Dudareva, N., Soltis, D.E., Hamberger, B., et al. (2019). A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. GigaScience **8**:giz005.

Zhao, Q., Chen, X.Y., and Martin, C. (2016a). *Scutellaria baicalensis*, the golden herb from the garden of Chinese medicinal plants. Sci. Bull. **61**:1391–1398.

Zhao, Q., Zhang, Y., Wang, G., Hill, L., Weng, J.K., Chen, X.Y., Xue, H., and Martin, C. (2016b). A specialized flavone biosynthetic pathway has evolved in the medicinal plant, *Scutellaria baicalensis*. Sci. Adv. **2**:e1501780.

Zhao, Q., Cui, M.Y., Levsh, O., Yang, D.F., Liu, J., Li, J., Hill, L., Yang, L., Hu, Y.H., Weng, J.K., et al. (2018). Two CYP82D enzymes function as flavone hydroxylases in the biosynthesis of root-specific 4'-deoxyflavones in *Scutellaria baicalensis*. Mol. Plant **11**:135–148.

Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., and Terry, J.M. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. **34**:303–311.