

1 **Comparative Genome Analysis of *Scutellaria baicalensis* and**  
2 ***Scutellaria barbata* Reveals the Evolution of Active Flavonoid**  
3 **Biosynthesis**

4

5 Zhichao Xu<sup>1,2,#</sup>, Ranran Gao<sup>1,#</sup>, Xiangdong Pu<sup>1</sup>, Rong Xu<sup>1,2</sup>, Jiyong Wang<sup>3</sup>, Sihao  
6 Zheng<sup>3</sup>, Yan Zeng<sup>3</sup>, Jun Chen<sup>1,2</sup>, Chunnian He<sup>1,2</sup>, Jingyuan Song<sup>1,2,\*</sup>

7

8 <sup>1</sup> *Key Lab of Chinese Medicine Resources Conservation, State Administration of*  
9 *Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal*  
10 *Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical*  
11 *College, Beijing 100193, China*

12 <sup>2</sup> *Engineering Research Center of Chinese Medicine Resource, Ministry of Education,*  
13 *Beijing 100193, China*

14 <sup>3</sup> *China National Traditional Chinese Medicine Co., Ltd, Beijing, 102600, China*

15

16 # Equal contribution.

17 \*Corresponding author(s).

18 E-mail: [jysong@implad.ac.cn](mailto:jysong@implad.ac.cn) (Song J)

19

20 Zhichao Xu: [zcxu@implad.ac.cn](mailto:zcxu@implad.ac.cn); 0000-0003-1753-5602

21 Ranran Gao: [grryzs@163.com](mailto:grryzs@163.com); 0000-0002-7164-0289

22 Xiangdong Pu: [xdpu\\_implad@163.com](mailto:xdpu_implad@163.com); 0000-0002-0892-035X

23 Rong Xu: [rxu@implad.ac.cn](mailto:rxu@implad.ac.cn); 0000-0002-0891-3686

24 Jiyong wang: [wangjiyong76@163.com](mailto:wangjiyong76@163.com); 0000-0002-0673-1803

25 Sihao Zheng: [zshzha@163.com](mailto:zshzha@163.com); 0000-0002-0523-7176

26 Yan Zeng: [zyzy1221@126.com](mailto:zyzy1221@126.com); 0000-0003-4128-5236

27 Jun Chen: [jchen@implad.ac.cn](mailto:jchen@implad.ac.cn); 0000-0001-5262-7444

28 Chunnian He: [cnhe@implad.ac.cn](mailto:cnhe@implad.ac.cn); 0000-0003-3659-7833

29 Jingyuan Song: [jysong@implad.ac.cn](mailto:jysong@implad.ac.cn); 0000-0003-2733-0416

30 **Running title:** *Xu Z et al / Genome of Scutellaria baicalensis and Scutellaria barbata*

31

32 Counts of words: 8390

33 Tables: 0

34 Figures: 4

35 Supplementary figures: 18

36 Supplementary tables: 19

37 Reference: 55

38 Reference from 2014: 40

39 **Abstract**

40 *Scutellaria baicalensis* and *Scutellaria barbata*, common medicinal plants of the  
41 Lamiaceae family, produce specific flavonoid compounds with antioxidant and  
42 antitumor activities, including baicalein, scutellarein, norwogonin, wogonin, and their  
43 glycosides. Here, we reported two chromosome-level genome assemblies of *S.*  
44 *baicalensis* and *S. barbata* with significant quantitative chromosomal variation ( $2n =$   
45  $18$  and  $2n = 26$ , respectively). The divergence of *S. baicalensis* and *S. barbata* occurred  
46 far earlier than previously reported, and a whole-genome duplication event was  
47 identified. The insertion of long terminal repeat elements after speciation might be  
48 responsible for the observed chromosomal expansion and rearrangement. The  
49 comparative genome analysis of congeneric species elucidated the species-specific  
50 evolution of chrysin and apigenin biosynthetic genes, such as the *S. baicalensis*-specific  
51 tandem duplication of the phenylalanine ammonia lyase (PAL) and chalcone synthase  
52 (CHS) genes, and the *S. barbata*-specific duplication of 4-CoA ligase (4CL) genes. In  
53 addition, the paralogous duplication, collinearity, and expression diversity of CYP82D  
54 subfamily members revealed the functional divergence of flavone hydroxylase genes  
55 between *S. baicalensis* and *S. barbata*. These *Scutellaria* genomes highlight the  
56 common and species-specific evolution of flavone biosynthetic genes, promoting the  
57 development of molecular breeding and the study of the biosynthesis and regulation of  
58 bioactive compounds.

59

60 **KEYWORDS:** *Scutellaria*; comparative genome; flavonoid biosynthesis; tandem  
61 duplication; species-specific evolution

62

63

## 64 **Introduction**

65 Plant-specific flavonoids, including flavones, flavonols, anthocyanins,  
66 proanthocyanidins, and isoflavones, play important functions in plants, such as flower  
67 pigmentation, UV protection, and symbiotic nitrogen fixation [1–3]. Flavonoid  
68 metabolites also have biological and pharmacological activities in human health,  
69 including antibacterial and antioxidant functions, and the treatment of cancer,  
70 inflammatory, and cardiovascular diseases [3]. The genus *Scutellaria*, belong to the  
71 Lamiaceae family, comprises common herbal plants enriched by bioactive flavonoids,  
72 and approximately 300 to 360 *Scutellaria* species have been documented as having the  
73 characteristic flower form of upper and lower lips [4,5]. Only two *Scutellaria* species,  
74 *Scutellaria baicalensis* and *Scutellaria barbata*, are recorded in the Chinese  
75 pharmacopoeia, and the roots of *S. baicalensis* and dried herbs of *S. barbata* are the  
76 basis of the Chinese medicines *Huang Qin* and *Ban Zhi Lian*, respectively, which have  
77 been well known heat-clearing and detoxifying herbs for thousands of years [6]. The  
78 main biologically active compounds in *Scutellaria* are derivatives of chrysin and  
79 apigenin, such as baicalein, scutellarein, wogonin, and their glycosides (baicalin,  
80 scutellarin, and wogonoside) [7–10]. Baicalin has been confirmed to activate carnitine  
81 palmitoyltransferase 1 in the treatment of diet-induced obesity and hepatic steatosis,  
82 leading to extensive interest in the potential antilipemic effect of this compound [11,12].

83 Illuminating the chemodiversity and biosynthesis of the active constituents of  
84 *Scutellaria* will provide a foundation for investigating the use of *Huang Qin* and *Ban*  
85 *Zhi Lian* in traditional Chinese medicine (TCM), and the production of these natural  
86 products via synthetic biology [13]. In *S. baicalensis*, the biosynthetic genes of the root-  
87 specific compounds baicalein and norwogonin have been functionally identified,  
88 providing an important basis for studying the biosynthesis and regulation of the natural  
89 products that make up *Huang Qin* [14,15]. Recently, the *in vitro* production of baicalein  
90 and scutellarein in *Escherichia coli* and *Saccharomyces cerevisiae* has been carried out  
91 based on the guidance of synthetic biology [16,17], but the metabolic engineering of  
92 these compounds still faces considerable challenges, including the discovery and

93 optimization of biological components. The *Salvia miltiorrhiza* genome from the  
94 Lamiaceae family provides useful information associated with secondary metabolism  
95 for the rapid functional identification of biosynthetic and regulatory genes [18–23]. In  
96 contrast, the genomes of the *Scutellara* genus remains unclear, and the reliance on  
97 transcriptome data from short-read sequencing has restricted gene discovery and  
98 analyses of genome evolution, including studies of gene family expansion and  
99 contraction, the evolution of biosynthetic genes, and identification of regulatory  
100 elements [24].

101 Significant morphological differences are present at the macroscopic level between  
102 *S. baicalensis* and *S. barbata*; these species are differentiation is mainly characterized  
103 by the fleshy rhizome and branched stem of *S. baicalensis* and the fibrous root and erect  
104 stem of *S. barbata*. The active compounds baicalein, wogonin and scutellarein are  
105 differentially distributed in the roots and aerial parts of *S. baicalensis* and *S. barbata*.  
106 Here, we performed *de novo* sequencing and assembly of the *S. baicalensis* and *S.*  
107 *barbata* genomes using a long-read strategy and Hi-C technology. The chromosome-  
108 level genome of *S. baicalensis* and *S. barbata* revealed their divergence time,  
109 chromosomal rearrangement and expansion, whole-genome duplication, and the  
110 evolutionary diversity of flavonoid biosynthesis. The study provided significant  
111 insights for the molecular assisted breeding of important TCM resources, genome  
112 editing, and understanding the molecular mechanisms of the chemodiversity of active  
113 compounds.

114

## 115 **Results and discussion**

### 116 **High-quality genome assemblies and annotation**

117 The size of the *S. baicalensis* genome was predicted to be  $440.2 \pm 10$  Mb and 441.9 Mb  
118 using flow cytometry and the 21 *k*-mer distribution analysis (approximately 0.96%  
119 heterozygosity) (**Figure 1A**, Figure S1). The genome survey of *S. barbata* showed a  
120 404.6 Mb genome size and 0.28% heterozygosity via the 21 *k*-mer distribution analysis  
121 (Figure 1A, Figure S1). Third-generation sequencing platforms, including PacBio and

122 Oxford Nanopore technologies, have been confirmed to have a significant advantage in  
123 *de novo* assembly and in processing data with complex structural variation due to high  
124 heterozygosity and repeat content [25–27]. Thus, 52.1 Gb Oxford Nanopore technology  
125 (ONT) reads ( $\sim 120 \times$ ) with an N50 of 16.3 kb from *S. baicalensis* and 51.7 Gb single  
126 molecule, real-time sequencing (SMRT) reads from the PacBio platform ( $\sim 130 \times$ ) with  
127 an N50 of 9.8 kb from *S. barbata* were produced to assemble highly contiguous  
128 genomes (Table S1). The low-quality long reads were further corrected and trimmed to  
129 yield 20.2 Gb ONT reads with an N50 of 35.5 kb from *S. baicalensis* and 18.0 Gb  
130 SMRT reads with an N50 of 15.3 kb from *S. barbata* using the CANU pipeline.

131 The contiguous assembly of the *S. baicalensis* and *S. barbata* genomes was  
132 performed using the optimized SMARTdenovo and  $3 \times$  Pilon polishing ( $50 \times$  Illumina  
133 reads) packages. For *S. baicalensis*, the contig-level genome assembly, which was  
134 377.0 Mb in length with an N50 of 2.1 Mb and a maximum contig length of 9.7 Mb  
135 covered 85.3% of the estimated genome size (Table S2). The *S. baicalensis* genome  
136 identified 91.5% of the complete Benchmarking Universal Single-Copy Orthologs  
137 (BUSCO) gene models and had an 88.7% DNA mapping rate, suggesting a high-quality  
138 genome assembly. For *S. barbata*, the contiguous contig assembly of 353.0 Mb with an  
139 N50 of 2.5 Mb and maximum contig of 10.5 Mb covered 87.2% of the predicted  
140 genome size (Table S2). The *S. barbata* genome identified 93.0% of complete BUSCO  
141 gene models and had a 95.0% DNA mapping rate. The high-quality genome assemblies  
142 of *S. baicalensis* and *S. barbata* showed the great advantage of single molecule  
143 sequencing, with assembly metrics that were far better than those of other reported  
144 genomes of Lamiaceae species, i.e., *Salvia miltiorrhiza* [28] and *Mentha longifolia* [29].

145 Given the assembly continuity, with a contig N50 of over 2 Mb for the *S. baicalensis*  
146 and *S. barbata* genomes, chromosome conformation capture (Hi-C) technology was  
147 applied to construct chromosome-level genomes [30]. In total, 99.8% and 98.8% of the  
148 assembled *S. baicalensis* and *S. barbata* contigs were corrected and anchored to 9 and  
149 13 pseudochromosomes ( $2n = 18$  for *S. baicalensis*,  $2n = 26$  for *S. barbata*) using a Hi-  
150 C interaction matrix with N50 values of 40.8 Mb and 23.7 Mb, respectively. The strong  
151 signal along the diagonal of interactions between proximal regions suggested that the

152 Hi-C assemblies for the *S. baicalensis* and *S. barbata* genomes had high quality (Figure  
153 S2).

154 The *S. baicalensis* genome comprised 33,414 protein-coding genes and 2,833  
155 noncoding RNAs (ncRNA), and 41,697 genes and 1,768 ncRNAs were annotated in  
156 the *S. barbata* genome (Table S4). Consistent with the genome assembly quality  
157 assessment, orthologs of 93.2% and 94.3% of the eukaryotic BUSCOs were identified  
158 in the *S. baicalensis* and *S. barbata* gene sets, suggesting the completeness of the  
159 genome annotation (Table S4). The gene-based synteny between *S. baicalensis* and *S.*  
160 *barbata* showed chromosome number variation and structural rearrangement (Figure  
161 1C, Figure S3, Table S3). In addition, the alignment at the DNA sequence level also  
162 showed the large-scale structural variations between *S. baicalensis* and *S. barbata*  
163 genome (Figure S4).

164

### 165 **Chromosome rearrangements and expansion after speciation**

166 Transposable elements (TEs) accounted for approximately 55.2% (208,004,279) and  
167 53.5% (188,790,851) of the *S. baicalensis* and *S. barbata* genomes, respectively (Table  
168 S5 and S6). And, 57.6% and 59.9% of these TEs were long terminal repeat (LTR)  
169 elements, respectively. Furthermore, we identified 1,225 and 1,654 full-length LTR  
170 elements, including *Gypsy* (342 and 310) and *Copia* (354 and 618) elements, in the *S.*  
171 *baicalensis* and *S. barbata* genomes (Table S7). However, there were significant  
172 differences in the insertion times of LTR elements, indicating that the LTRs (1.41 MYA,  
173 million years ago) in *S. baicalensis* are more ancient than those (0.88 MYA) in *S.*  
174 *barbata*, assuming a mutation rate of  $\mu=1.3\times 10^{-8}$  (per bp per year) (Figure S5, Table  
175 S7). The recent insertion and activation of LTRs might be key factors in the generation  
176 of chromosome rearrangements and expansion of *S. barbata* [31,32]. The ribosomal  
177 RNAs (rRNAs) and simple sequence repeats (SSRs) were further annotated (Table S8  
178 and S9). A total of 142,951 and 147,705 SSRs were annotated in *S. baicalensis* and *S.*  
179 *barbata*, respectively, and these SSRs will provide useful molecular markers for  
180 breeding and genetic diversity studies.

181 We employed a genome-wide high-resolution Hi-C interaction analysis of *S.*

182 *baicalensis* and *S. barbata* to characterize the architectural features of folded eukaryotic  
183 chromatin, including interchromosomal interactions, the compendium of chromosomal  
184 territories, and A/B compartments [33–35]. First,  $159 \times$  and  $173 \times$  Hi-C sequencing  
185 reads were uniquely mapped (49.6% and 59.0%) to the *S. baicalensis* and *S. barbata*  
186 reference genomes, respectively. Then, 84.8 and 113.1 million valid interaction pairs  
187 were obtained to construct the matrix of interactions among 100 kb binned genomic  
188 regions across all 9 *S. baicalensis* chromosomes and 13 *S. barbata* chromosomes. The  
189 whole-chromosome interactions of *S. baicalensis* indicated that chr5 and chr9 had a  
190 closer association than the other chromosome pairs. In *S. baicalensis*, the chromosome  
191 set including chr2, chr3 and chr8 showed enrichment and association with each other,  
192 and depletion with other interchromosomal sets, implying that these three  
193 chromosomes were mutually closer in space than the other chromosomes (Figure S6).  
194 In *S. barbata*, the chromosomal territories of chr4, chr5, and chr9, with mutual  
195 interactions, occupied an independent region in the nucleus (Figure S7).

196 As the secondary major structural unit of chromatin packing in *S. baicalensis* and  
197 *S. barbata*, the A/B compartments representing open and closed chromatin, respectively,  
198 were characterized according to an eigenvector analysis of the genome contact matrix.  
199 Similarly, more than half of the assembled *S. baicalensis* and *S. barbata* genomes (53.2%  
200 and 52.0%) were identified as A compartment in the leaf tissue. As expected, the TE  
201 density in the A compartment was dramatically lower than that in the B compartment  
202 ( $p < 0.001$ ), and the gene number per 100 kb was significantly higher in the A  
203 compartment ( $p < 0.001$ ) (Figure S5 and S6), indicating a positive correlation between  
204 the A compartment and transcriptional activity or other functional measures [33,35].

205

#### 206 **Whole-genome duplication events between *S. baicalensis* and *S. barbata***

207 Conserved sequences, including orthologs and paralogs, can be used to deduce  
208 evolutionary history based on whole-genome comparisons. Here, orthologous groups  
209 of amino acid sequences from 11 angiosperms were identified, yielding a total of 19,479  
210 orthologous groups that covered 291,192 genes. Among these, 120,459 genes clustering  
211 into 6,837 groups were conserved in all examined plants. Computational analysis of



212 gene family evolution (CAFÉ) showed that 1,180 and 1,853 gene families were  
213 expanded in the *S. baicalensis* and *S. barbata* lineages, respectively, while 1,599 and  
214 1,632 gene families contracted, respectively (Figure 2A, Figure S8, Table S10).  
215 Functional exploration of *Scutellaria*-specific genes indicated that domains related to  
216 secondary metabolite biosynthesis, such as transcription factors, cytochrome P450s,  
217 and O-methyltransferase were markedly enriched.

218 In addition, 235 single-copy genes in all tested plants were identified and used to  
219 construct a phylogenetic tree, indicating that these two *Scutellaria* species were most  
220 closely related to *Salvia miltiorrhiza* with an estimated divergence time of 41.01 MYA;  
221 *S. baicalensis* and *S. barbata* were grouped into one branch, with an estimated  
222 divergence time of approximately 13.28 MYA (Figure 2A). The Phylogenetic tree also  
223 supported the close relationship between Lamiaceae (*S. baicalensis*, *S. barbata* and *S.*  
224 *miltiorrhiza*) and Pedaliaceae (*Sesamum indicum*) with the divergence time of  
225 approximately 49.90 MYA (Figure 2A) [36]. Previous research reported that the  
226 divergence time of *S. baicalensis* and *S. barbata* based on the *matK* and *CHS* (chalcone  
227 synthase) genes was ~3.35 MYA [37]. However, a genome-wide analysis identified 8  
228 and 3 *CHS* genes in *S. baicalensis* and *S. barbata*, respectively, and the expansion and  
229 evolution of *CHS* negatively impacted the estimation of diversification history between  
230 these *Scutellaria* species.

231 Based on sequence homology, 17,265 orthologous gene pairs with synteny were  
232 identified between the *S. baicalensis* and *S. barbata* genomes, and the distribution of  
233 synonymous substitution rates (*Ks*) peaked at approximately 0.16, representing the  
234 speciation time of *S. baicalensis* and *S. barbata* (Figure 2B, Table S11). The mean *Ks*  
235 values of orthologous gene pairs with synteny and the divergence times among *S.*  
236 *baicalensis*, *S. barbata*, *S. miltiorrhiza*, *S. indicum*, and *Vitis vinifera* [38], showed the  
237 estimated synonymous substitutions per site per year as  $1.30 \times 10^{-8}$  for the test species  
238 (Table S11). In total, 7,812, 7,168, 6,984, and 7,711 paralogous gene pairs were  
239 identified, and the distribution of *Ks* values peaked at approximately 0.87, 0.86, 1.02  
240 and 0.67 in *S. baicalensis*, *S. barbata*, *S. miltiorrhiza* and *S. indicum*, respectively  
241 (Figure 2B, Table S11). Based on the phylogenetic analysis, the WGD event happened

242 before the divergence of *S. baicalensis*, *S. barbata*, *S. miltiorrhiza* and *S. indicum*. Then,  
243 we traced the divergence time of Lamiaceae and Pedaliaceae shared WGD event around  
244 46.24-60.71 MYA (Table S11). The distribution of the  $K_s$  values of paralogous genes  
245 showed that no whole-genome duplication (WGD) events have occurred since the  
246 divergence of *S. miltiorrhiza*, *S. baicalensis* and *S. barbata*. Comparison of *S.*  
247 *baicalensis* and *S. barbata* genomes with an ancestral eudicot karyotype (AEK) genome  
248 [39], and with grape genome, also supported the structural rearrangement between *S.*  
249 *baicalensis* and *S. barbata* genomes, and the shared WGD event after WGT- $\gamma$  event of  
250 grape (Figure 2C, Figure S9). The genome syntenic analysis indicated two copies of  
251 syntenic blocks from of Lamiaceae and Pedaliaceae species per corresponding grape  
252 block, which confirmed the recent WGD event before the divergence of *S. baicalensis*,  
253 *S. barbata*, *S. indicum* (Figure S10).

254

### 255 **Organ-specific localization of bioactive compounds**

256 Baicalein, scutellarein, norwogonin, wogonin, and their glycosides (baicalin,  
257 scutellarin, norwogonoside and wogonoside) are the main bioactive compounds in *S.*  
258 *baicalensis* and *S. barbata*. We collected samples from the root, stem, leaf and flower  
259 tissues of *S. baicalensis* and *S. barbata* to detect the accumulation of active compounds.  
260 The results indicated that baicalein, norwogonin, wogonin, baicalin, norwogonoside  
261 and wogonoside mainly accumulated in the roots of *S. baicalensis* and *S. barbata*, while  
262 scutellarin was distributed in the aerial parts (stem, leaf and flower) of these species  
263 (Figure 1B, Figure S11, Table S12), providing a potential basis for the co-expression  
264 analysis of biosynthetic genes [23].

265 Transcriptome analysis of these four tissues from *S. baicalensis* and *S. barbata*  
266 included calculation of the FPKM values of 39,121 and 47,200 genes, respectively.  
267 Among them, 31.5% (12,320) and 40.6% (19,153) of the transcripts were not expressed  
268 (FPKM < 1) in any of the tested tissues. Based on k-means clustering, all the expressed  
269 genes from *S. baicalensis* and *S. barbata* were clustered into 48 groups (Figure S12 and  
270 S13). The expression levels of 3,421 genes from clusters 8, 20, 32, 33, 34, 39, and 47  
271 in *S. baicalensis*, and 3,675 genes from clusters 2, 4, 21, 25, 27, 31, and 40 in *S. barbata*

272 were significantly higher in the roots than in the other organs. The biosynthetic genes  
273 involved in the synthesis of *Scutellaria* specific flavones and glycosides, containing  
274 genes encoding chalcone synthase, chalcone isomerase, CYP450s, O-methyltransferase,  
275 glycosyltransferase and glycosyl hydrolases, were enriched, with high expression in the  
276 roots of *S. baicalensis* and *S. barbata* (Table S13 and S14).

277

### 278 **Conserved evolution of the chrysin and apigenin biosynthetic pathways in *S.*** 279 ***baicalensis* and *S. barbata***

280 The main active compounds in the medicinal plants *S. baicalensis* and *S. barbata* are  
281 flavonoids, and the chrysin biosynthetic genes in *S. baicalensis* encoding 4-CoA ligase  
282 (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), and flavone synthase  
283 (FNSII) have been cloned and functionally identified [14]. However, the gene locations,  
284 gene numbers and evolution of this pathway in the *S. baicalensis* and *S. barbata*  
285 genomes remain unclear. Here, we identified the same number of chrysin and apigenin  
286 biosynthetic genes in the *S. baicalensis* and *S. barbata* genomes and determined the  
287 expression levels of these genes, including phenylalanine ammonia lyase (PAL, 5 and  
288 4), cinnamate 4-hydroxylase (C4H, 3 and 4), 4CL (9 and 14), CHS (8 and 3), CHI (1  
289 and 1), and FNSII (3 and 3), in different tissues (**Figure 3A**, Table S15 and S16). Here,  
290 18 orthologous gene pairs were found between the *S. baicalensis* and *S. barbata*  
291 genomes, and the *Ka/Ks* value (average 0.13) indicated purifying selection on flavone  
292 biosynthesis during evolution [40] (Figure 3B, Table S17). The PAL and CHS gene  
293 numbers in *S. baicalensis* were expanded compared to those in *S. barbata*; conversely,  
294 a significant duplication event of 4CL genes in *S. barbata* was found, suggesting that  
295 expansion via tandem duplication might have occurred after the separation of these  
296 *Scutellaria* species. The *Ks* values of 18 orthologous gene pairs of *S. baicalensis* and *S.*  
297 *barbata* in the chrysin and apigenin biosynthetic pathways indicated that the specific  
298 expansion of the SbaiPAL (SbaiPAL1 and SbaiPAL2), SbaiCHS (SbaiCHS2, SbaiCHS3,  
299 SbaiCHS4, and SbaiCHS5) and Sbar4CL (Sbar4CL1-1 and Sbar4CL1-2, Sbar4CL1-3  
300 and Sbar4CL1-4, Sbar4CLL9-2 and Sbar4CLL9-3) genes had occurred via tandem

301 duplication, after the speciation of *S. baicalensis* and *S. barbata* (Figure 3, Figure S14,  
302 Table S17).

303 *Sbai4CLL7* and *SbaiCHS1* have been reported to be related to the biosynthesis of  
304 specific 4'-deoxyflavones with cinnamic acid as a substrate in *S. baicalensis* [14].  
305 Compared to *S. miltiorrhiza*, the 4CLL7 genes from the *Scutellaria* genus showed gene  
306 expansion, and the gene duplication of *Sbai4CLL7-1* and *Sbai4CLL7-2* occurred before  
307 the speciation of *S. baicalensis* and *S. barbata* (Figure S13). *Sbai4CLL7-1* and  
308 *Sbar4CLL7-1* showed no expression in the tested transcriptomes, and the duplication  
309 of the *Scutellaria*-specific 4CLL7-2 allowed the evolution of substrate preferences for  
310 the catalysis of cinnamic acid. The initial step and central hub for flavone biosynthesis  
311 is the catalysis of CHS; hence, the expression of CHS is required for the production of  
312 flavonoids, isoflavonoids, and other metabolites in plants [41]. Here, we also detected  
313 the highest expression levels of *SbaiCHS1* and *SbarCHS1* in all the tested samples;  
314 however, a recent expansion of CHS genes has occurred in *S. baicalensis*, and 4  
315 additional paralogs of *SbaiCHS1* (*Sbai7C107T21*) were observed in chr7. Duplications  
316 of the *SbaiCHS2*, *SbaiCHS3*, *SbaiCHS4* and *SbaiCHS5* genes occurred after the  
317 speciation of *S. baicalensis* and *S. barbata* (Figure 3C). The nucleotide and amino acid  
318 sequences of *SbaiCHS2* and *SbaiCHS3* were identical, but *SbaiCHS5* contained a  
319 variant K316 deletion. The divergence of *SbaiCHS1* and *SbarCHS1* occurred before the  
320 separation of *S. miltiorrhiza* and the *Scutellaria* species, suggesting a conserved  
321 function of chalcone synthase in flavone biosynthesis. In addition, the tandemly  
322 duplicated *SbaiCHS2-5* genes were more highly expressed in the roots of *S. baicalensis*  
323 than in other tissues (Figure 3C), suggesting that their species-specific evolution might  
324 be related to the biosynthesis of flavones and their glycosides, which are enriched in  
325 roots.

326 C4H is responsible for the biosynthesis of coumaroyl-CoA, which might be the  
327 restrictive precursor of the 4'-hydroxyl group involved in scutellarein biosynthesis.  
328 Here, we identified high expression of *SbaiC4H1* and *SbarC4H1* in the stems, leaves,  
329 and flowers of *S. baicalensis* and *S. barbata* (Figure 3B, Figure S14). This high

330 expression level was positively correlated with the distribution of scutellarein, which is  
331 biosynthesized in the aerial parts of *S. baicalensis* and *S. barbata* (Figure 1B).

332 The *SbaiFNSII2* gene, which has been reported to catalyze the formation of chrysin  
333 in *S. baicalensis*, presented high expression in the roots and stems, and its ortholog  
334 *SbarFNSII2* was also significantly expressed in the roots of *S. barbata*. A genome  
335 collinearity analysis identified 566 orthologous gene pairs covering a region ~6 Mb in  
336 length across chr3 of *S. baicalensis* and chr13 of *S. barbata*, including the tandem  
337 duplication of *SbaiFNSII1-SbaiFNSII2* and *SbarFNSII1-SbarFNSII2*. This duplication  
338 occurred before the speciation of *S. baicalensis* and *S. barbata* (Figure S14). The  
339 majority of the FNSII region (~85%) in *S. baicalensis* and *S. barbata* was assigned to  
340 the A compartment, indicating high transcriptional activity. The genome synteny of the  
341 FNSII region between *S. baicalensis* and *S. barbata* suggested the conserved evolution  
342 of flavone synthase.

343

#### 344 **Functional divergence of flavone hydroxylase genes between *S. baicalensis* and *S.*** 345 ***barbata***

346 CYP450 superfamily members, such as C4H (CYP73A family), FNSII (CYP93B  
347 family), flavone 6-hydroxylase (F6H, CYP82D family) and flavone 8-hydroxylase  
348 (F8H, CYP82D family), perform key modifications in flavone biosynthesis.  
349 *SbaiCYP82D1* has been reported to have 6-hydroxylase activity on chrysin and  
350 apigenin to produce baicalein and scutellarein, respectively, and *SbaiCYP82D2* can  
351 catalyze chrysin to norwogonin in *S. baicalensis* [15] (Figure S15). Here, we identified  
352 418 and 398 CYP450 gene members, and 17 and 24 physical clusters of CYP450s (5  
353 gene clusters per 500 kb) in the *S. baicalensis* and *S. barbata* genomes, respectively  
354 (Figure S16 and S17), suggesting a high frequency of CYP gene tandem duplication.  
355 Among them, 18 CYP82D members containing *SbaiCYP82D1-9* and *SbarCYP82D1-*  
356 *9* were identified in the *S. baicalensis* and *S. barbata* genomes; these genes might be  
357 responsible for the hydroxylation of chrysin and apigenin (Table S18). Consistent with  
358 a previous report, significant expression of *SbaiCYP82D1* and *SbaiCYP82D2* in the  
359 roots of *S. baicalensis* was detected, in accordance with the accumulation of baicalein,

360 wogonin, and their glycosides (**Figure 4A**). However, *SbarCYP82D1* showed relatively  
361 high expression in stems and leaves, and *SbarCYP82D2* showed extremely low  
362 expression in all tissues of *S. barbata*, in contrast to the distributions of active flavones,  
363 suggesting a potential functional divergence of hydroxylation between *S. baicalensis*  
364 and *S. barbata*.

365 Three-gene tandem duplications of *SbaiCYP82D1-SbaiCYP82D7-SbaiCYP82D8*  
366 and *SbarCYP82D1-SbarCYP82D6-SbarCYP82D8* (physical distance < 30 kb) on chr6  
367 of *S. baicalensis* and *S. barbata* were identified (Figure 4B). According to the 150 kb  
368 collinearity analysis, 11 orthologous gene pairs, including CYP82D8 from *S.*  
369 *baicalensis* and *S. barbata*, presented conserved evolution. The phylogenetic analysis  
370 and *Ks* values of orthologous gene pairs indicated that the duplication of *SbarCYP82D8*  
371 and *SbarCYP82D6* occurred after the speciation of *S. barbata* (Table S19); however,  
372 duplication of *SbaiCYP82D8* and *SbaiCYP82D7* happened before the divergence of *S.*  
373 *baicalensis* and *S. barbata* (Figure 4D, Figure S18). This contradiction and  
374 evolutionary divergence supports the following proposed hypothesis: 1) the first  
375 duplication of *CYP82D8* produced the new *CYP82D1*, and the duplication event  
376 occurred around WGD event. 2) the second duplication of *CYP82D8* generated the new  
377 *CYP82D7*, similar to the tandem duplication of *SbaiCYP82D8-SbaiCYP82D7-*  
378 *SbaiCYP82D1* in *S. baicalensis*. 3) After speciation, the third duplication event of  
379 *SbarCYP82D8* uniquely occurred in the *S. barbata* genome and produced  
380 *SbarCYP82D6*; a recent gene transfer of *SbarCYP82D7* via transposon from chr6 to  
381 chr3 in *S. barbata* was predicted. An adjacent intact LTR/ Gypsy in *SbarCYP82D7* was  
382 identified, and its insertion time was estimated to be ~3.5 MYA. Given the evolution  
383 and high expression of *SbarCYP82D6* and *SbarCYP82D8*, we speculated that these two  
384 genes might be responsible for the F6H function in chrysin and apigenin synthesis *in*  
385 *vivo* in *S. barbata*.

386 The chromosome location of F8H functional members showed that *SbaiCYP82D2*,  
387 *SbaiCYP82D3*, *SbaiCYP82D4*, *SbaiCYP82D5*, *SbaiCYP82D6* and *SbaiCYP82D9* were  
388 distributed on chr1 of *S. baicalensis*, and *SbarCYP82D2*, *SbarCYP82D3*,  
389 *SbarCYP82D4*, *SbarCYP82D5* and *SbarCYP82D9* were located on chr7 of *S. barbata*.

390 The structural rearrangement of large segments between chr1 of *S. baicalensis* and chr7  
391 of *S. barbata* was found (Figure 4C, Figure S4). In addition, tandem duplications  
392 containing three CYP genes (*SbaiCYP82D2-SbaiCYP82D3-SbaiCYP82D5* and  
393 *SbarCYP82D3-SbarCYP82D2-SbarCYP82D4*) were identified (Figure 4C). The  
394 orthologous gene pairs (*SbaiCYP82D2-SbarCYP82D2* and *SbaiCYP82D3-*  
395 *SbarCYP82D3*) presented high identity values of 90.11% and 83.72%. The duplications  
396 of *SbarCYP82D3-SbarCYP82D4*, *SbaiCYP82D4-SbaiCYP82D5*, and *SbaiCYP82D6-*  
397 *SbaiCYP82D9* occurred after the speciation of *S. baicalensis* and *S. barbata* (Table S19).  
398 However, the expression of *SbarCYP82D2*, *SbarCYP82D3* and *SbarCYP82D4* is slight  
399 in *S. barbata*, indicating functional divergence following species-specific duplication  
400 events. In contrast, the *SbarCYP82D5* and *SbarCYP82D9* were highly expressed in the  
401 roots of *S. barbata*, suggesting a potential F8H function in the biosynthesis of  
402 norwogonin.

403

## 404 **Conclusions**

405 We reported two chromosome-level genomes of the medicinal plants *S. baicalensis* and  
406 *S. barbata* through the combination of second-generation sequencing (Illumina  
407 platform), third-generation sequencing (PacBio and Oxford Nanopore platforms), and  
408 Hi-C technologies. This study confirmed and traced the divergence times of *S.*  
409 *baicalensis* and *S. barbata*, which occurred 13.28 MYA, far earlier than previously  
410 reported. Comparative genomic analysis revealed similar TE proportions in the *S.*  
411 *baicalensis* and *S. barbata* genomes, while the recent LTR insertion in *S. barbata* might  
412 be an important factor resulting in chromosomal rearrangement and expansion. A WGD  
413 event (~52.11-78.84 MYA ) shared among *S. baicalensis*, *S. barbata*, *S. miltiorrhiza*,  
414 and *S. indicum*. The tandem duplication of paralogs after the speciation of *S. baicalensis*  
415 and *S. barbata* might be the most important contributor to the divergent evolution of  
416 flavonoid biosynthetic gene families, such as PAL, 4CL CHS, F6H and F8H. A  
417 determination of the distribution of flavone contents and transcriptome analysis  
418 supported the functional divergence of flavonoid biosynthetic genes between *S.*

419 *baicalensis* and *S. barbata*. The two high-quality genomes reported in the present study  
420 will enrich genome research in the Lamiaceae and provide important insights for studies  
421 of breeding, evolution, chemodiversity and genome editing.

422

## 423 **Materials and methods**

### 424 **Plant materials**

425 *S. baicalensis* and *S. barbata* plants were cultivated in the experimental field of the  
426 IMPLAD (Institute of Medicinal Plant Development) (40°N and 116°E), Beijing, China.  
427 Four independent tissues from *S. baicalensis* and *S. barbata*, namely, root, stem, leaf,  
428 and flower tissues, were collected in three replicates. These tissues were used separately  
429 for the measurement of active compounds and RNA sequencing. High-quality DNA  
430 extracted from young leaves was used to construct libraries for Illumina, ONT and  
431 Sequel sequencing.

432

### 433 **Long-read sequencing and assemblies**

434 The high-molecular-weight (HMW) genomic DNA of *S. baicalensis* and *S. barbata* was  
435 extracted in accordance with the method for megabase-sized DNA preparation [42].  
436 HMW gDNA fragments (>20 kb) were selected using BluePippin. Long-read libraries  
437 were constructed following the protocols for the ONT (<https://nanoporetech.com/>) and  
438 PacBio Sequel platforms (<https://www.pacb.com/>). The ONT reads of *S. baicalensis*  
439 were generated using the ONT GridION X5 platform, and the library of *S. barbata* was  
440 sequenced using the Sequel platform. The raw ONT and SMRT reads were filtered via  
441 MinKNOW and SMRT Link, respectively. First, CANU (v1.7) was used to correct and  
442 trim the long reads from the ONT and Sequel platforms with the default parameters  
443 [43]. Then, the corrected and trimmed ONT and SMRT reads were assembled using  
444 SMARTdenovo (<https://github.com/ruanjue/smartdenovo>). Finally, Illumina short  
445 reads were used to polish the assembled contigs three times using Pilon (v1.22). The  
446 quality of the genome assemblies was estimated by a BUSCO (v2.0) search [44] and  
447 by mapping Illumina reads from the DNA and RNA libraries to the assembled genomes.



448

#### 449 **Chromosome construction using Hi-C**

450 Young leaves from *S. baicalensis* and *S. barbata* were fixed and crosslinked, and then,  
451 Hi-C libraries were constructed and sequenced using Illumina as described [33,34]. The  
452 short reads were mapped to the assembled genome using BWA [45], and the valid  
453 interaction pairs were selected using Hi-C Pro [46]. Then, the draft assemblies of *S.*  
454 *baicalensis* and *S. barbata* were anchored to chromosomes ( $2n = 18$  and  $2n = 26$ ,  
455 respectively) using LACHESIS with the following parameters: CLUSTER MIN RE  
456 SITES = 62, CLUSTER MAX LINK DENSITY = 2, CLUSTER NONINFORMATIVE  
457 RATIO = 2, ORDER MIN N RES IN TRUN = 53, ORDER MIN N RES IN SHREDS  
458 = 52 [30].

459

#### 460 **Genome annotation**

461 The RepeatModeler (v1.0.9) package, including RECON and RepeatScout, was used  
462 to identify and classify the repeat elements of the *S. baicalensis* and *S. barbata* genomes.  
463 The repeat elements were then masked by RepeatMasker (v4.0.6). The long terminal  
464 repeat retrotransposons (LTR-RTs) in *S. baicalensis* and *S. barbata* were identified  
465 using LTR\_Finder (v1.0.6) and LTR\_retriever. Twenty-four samples from a total of  
466 eight different *S. baicalensis* and *S. barbata* tissues (roots, stems, leaves, and flowers)  
467 were subjected to RNA-Seq using the Illumina NovaSeq platform. The clean reads from  
468 *S. baicalensis* and *S. barbata* were *de novo* assembled using Trinity (v 2.2.0), and the  
469 coding regions in the assembled transcripts were predicted using TransDecoder (v2.1.0).  
470 The gene annotation of the masked *S. baicalensis* and *S. barbata* genome was *ab initio*  
471 predicted using the MAKER (v2.31.9) pipeline, integrating the assembled transcripts  
472 and protein sequences from *S. baicalensis*, *S. barbata*, and *A. thaliana* [47]. Noncoding  
473 RNAs and miRNAs were annotated by alignment to the Rfam and miRNA databases  
474 using INFERNAL (v1.1.2) and BLASTN, respectively. RNA-Seq reads from different  
475 *S. baicalensis* and *S. barbata* tissues were mapped to the masked genome using HISAT2  
476 (v2.0.5), and the different expression levels of the annotated genes were calculated  
477 using Cufflinks (v2.2.1) [48].

478

## 479 **Genome evolution analysis**

480 The full amino acid sequences of *S. baicalensis*, *S. barbata* and nine other angiosperms  
481 were aligned to orthologous groups using OrthoFinder [49]. The basal angiosperm  
482 *Amborella trichopoda*, was chosen as the outgroup. Single-copy genes were used to  
483 construct a phylogenetic tree using the RAxML package with PROTGAMMAJTT  
484 model and 1000 replicates (version 8.1.13). The divergence times among 11 plants were  
485 predicted using r8s program based on the estimated divergence times  
486 *Amborella trichopoda-Vitis vinifera* (173-199 MYA) and *Populus trichocarpa*-  
487 *Arabidopsis thaliana* (98-117 MYA). According to the phylogenetic analysis and  
488 divergence times, expansion and contraction of the gene families were identified using  
489 CAFÉ (v 3.1) [50]. The paralogous and orthologous gene pairs from *S. baicalensis*, *S.*  
490 *barbata*, and *S. miltiorrhiza* were identified, and the *Ka*, *Ks* and *Ka/Ks* values of *S.*  
491 *baicalensis-S. baicalensis*, *S. barbata-S. barbata*, *S. miltiorrhiza-S. miltiorrhiza*, *S.*  
492 *baicalensis-S. miltiorrhiza*, *S. baicalensis-S. barbata*, and *S. barbata-S. miltiorrhiza*,  
493 were calculated using the SynMap2 and DAGchainer method of CoGE Comparative  
494 Genomics Platform. The detection of synteny and collinearity among *S. baicalensis*, *S.*  
495 *barbata*, and *S. miltiorrhiza* was performed using MCscan X(v1.1) [51].

496

## 497 **Identification of gene families related to flavone biosynthesis**

498 Protein sequences of the PAL, 4CL, C4H, CHS, CHI, and FNSII gene family members  
499 in *A. thaliana* were downloaded from the TAIR database, and F6H and F8H in *S.*  
500 *baicalensis* were obtained from a previous study. Then, these sequences were searched  
501 against the *S. baicalensis* and *S. barbata* protein sequences using BLASTP with an E  
502 value cutoff of 1e-10. The conserved domains of the protein sequences of candidate  
503 genes were further searched in the Pfam database using hidden Markov models [52].  
504 Full-length protein sequences were used to construct phylogenetic trees using the  
505 maximum likelihood method with the Jones-Taylor-Thornton (JTT) model and 1,000  
506 bootstrap replicates [53]. A detailed description of some materials and methods used is  
507 provided in Supplementary methods and results.

508

### 509 **Data availability**

510 The raw sequence data reported in this paper have been deposited in the Genome  
511 Sequence Archive [54] in BIG Data Center [55], Beijing Institute of Genomics (BIG),  
512 Chinese Academy of Sciences, under accession numbers CRA001730 that are publicly  
513 accessible at <http://bigd.big.ac.cn/gsa>. The assembled genomes and gene structures  
514 were also submitted to CoGe with id54175 for *S. baicalensis* and id54176 for *S. barbata*.

515

### 516 **Authors' Contributions**

517 ZX and JS designed and coordinated the study. ZX assembled and analyzed the genome.  
518 RX, JW, SZ, YZ, and JC supplied plant materials. RG, XP, and CH performed the  
519 experiments and analyzed the data. ZX, RG, and JS wrote and edited the manuscript.

520

### 521 **Competing interests**

522 The authors have declared no competing interests.

523

### 524 **Acknowledgments**

525 This work was supported by the National Natural Science Foundation of China (Grant  
526 No. 31700264) and the Chinese Academy of Medical Sciences (CAMS) Innovation  
527 Fund for Medical Sciences (CIFMS) (Grant No. 2016-I2M-3-016).

528

529 **References:**

530

531 [1] Winkel-Shirley B. Flavonoid biosynthesis. A colorful model for genetics,  
532 biochemistry, cell biology, and biotechnology. *Plant Physiol* 2001; 126: 485–93.

533 [2] Winkel-Shirley B. Biosynthesis of flavonoids and effects of stress. *Curr Opin Plant*  
534 *Biol* 2002; 5: 218–23.

535 [3] Grotewold E. The genetics and biochemistry of floral pigments. *Annu Rev Plant*  
536 *Biol* 2006; 57: 761–80.

537 [4] Shang X, He X, He X, Li M, Zhang R, Fan P, et al. The genus *Scutellaria* an  
538 ethnopharmacological and phytochemical review. *J Ethnopharmacol* 2010; 128: 279–  
539 313.

540 [5] Grzegorzcyk-Karolak I, Wiktorek-Smagur A, Hnatuszko-Konka K. An untapped  
541 resource in the spotlight of medicinal biotechnology: the genus *Scutellaria*. *Curr Pharm*  
542 *Biotechnol* 2018; 19: 358–71.

543 [6] Chinese Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of  
544 China. Beijing: China Medical Science Press; 2015.

545 [7] Zhang Z, Lian XY, Li S, Stringer JL. Characterization of chemical ingredients and  
546 anticonvulsant activity of American skullcap (*Scutellaria lateriflora*). *Phytomedicine*  
547 2009; 16: 485–93.

548 [8] Qiao X, Li R, Song W, Miao WJ, Liu J, Chen HB, et al. A targeted strategy to analyze  
549 untargeted mass spectral data: Rapid chemical profiling of *Scutellaria baicalensis* using  
550 ultra-high performance liquid chromatography coupled with hybrid quadrupole orbitrap

- 551 mass spectrometry and key ion filtering. *J Chromatogr A* 2016; 1441: 83–95.
- 552 [9] Yan B, Xu W, Su S, Zhu S, Zhu Z, Zeng H, et al. Comparative analysis of 15  
553 chemical constituents in *Scutellaria baicalensis* stem-leaf from different regions in  
554 China by ultra-high performance liquid chromatography with triple quadrupole tandem  
555 mass spectrometry. *J Sep Sci* 2017; 40: 3570–81.
- 556 [10] Zhao Q, Chen XY, Martin C. *Scutellaria baicalensis*, the golden herb from the  
557 garden of Chinese medicinal plants. *Sci Bull (Beijing)* 2016; 61: 1391–8.
- 558 [11] Dai J, Liang K, Zhao S, Jia W, Liu Y, Wu H, et al. Chemoproteomics reveals  
559 baicalin activates hepatic CPT1 to ameliorate diet-induced obesity and hepatic steatosis.  
560 *Proc Natl Acad Sci U S A* 2018; 115: E5896–905.
- 561 [12] Guo HX, Liu DH, Ma Y, Liu JF, Wang Y, Du ZY, et al. Long-term baicalin  
562 administration ameliorates metabolic disorders and hepatic steatosis in rats given a  
563 high-fat diet. *Acta Pharmacol Sin* 2009; 30: 1505–12.
- 564 [13] Chen S, Song J, Sun C, Xu J, Zhu Y, Verpoorte R, et al. Herbal genomics:  
565 examining the biology of traditional medicines. *Science* 2015; 347: S27–8.
- 566 [14] Zhao Q, Zhang Y, Wang G, Hill L, Weng JK, Chen XY, et al. A specialized flavone  
567 biosynthetic pathway has evolved in the medicinal plant, *Scutellaria baicalensis*. *Sci*  
568 *Adv* 2016; 2: e1501780.
- 569 [15] Zhao Q, Cui MY, Levsh O, Yang D, Liu J, Li J, et al. Two CYP82D enzymes  
570 function as flavone hydroxylases in the biosynthesis of root-specific 4'-deoxyflavones  
571 in *Scutellaria baicalensis*. *Mol Plant* 2018; 11: 135–48.
- 572 [16] Liu X, Cheng J, Zhang G, Ding W, Duan L, Yang J, et al. Engineering yeast for the

- 573 production of breviscapine by genomic analysis and synthetic biology approaches. Nat  
574 Commun 2018; 9: 448.
- 575 [17] Li JH, Tian CF, Xia YH, Mutanda I, Wang KB, Wang Y. Production of plant-  
576 specific flavones baicalein and scutellarein in an engineered *E. coli* from available  
577 phenylalanine and tyrosine. Metab Eng 2019; 52: 124–33.
- 578 [18] Xu Z, Song J. The 2-oxoglutarate-dependent dioxygenase superfamily participates  
579 in tanshinone production in *Salvia miltiorrhiza*. J Exp Bot 2017; 68: 2299–308.
- 580 [19] Cao W, Wang Y, Shi M, Hao X, Zhao W, Wang Y, et al. Transcription factor  
581 SmWRKY1 positively promotes the biosynthesis of tanshinones in *Salvia miltiorrhiza*.  
582 Front Plant Sci 2018; 9: 554.
- 583 [20] Huang Q, Sun M, Yuan T, Wang Y, Shi M, Lu S, et al. The AP2/ERF transcription  
584 factor SmERF1L1 regulates the biosynthesis of tanshinones and phenolic acids in  
585 *Salvia miltiorrhiza*. Food Chem 2019; 274: 368–75.
- 586 [21] Sun M, Shi M, Wang Y, Huang Q, Yuan T, Wang Q, et al. The biosynthesis of  
587 phenolic acids is positively regulated by the JA-responsive transcription factor ERF115  
588 in *Salvia miltiorrhiza*. J Exp Bot 2019; 70: 243–54.
- 589 [22] Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the genome sequence  
590 of the medicinal plant *Salvia miltiorrhiza*. Mol Plant 2016; 9: 949–52.
- 591 [23] Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, et al. Full-length  
592 transcriptome sequences and splice variants obtained by a combination of sequencing  
593 platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone  
594 biosynthesis. Plant J 2015; 82: 951–61.

- 595 [24] Xin T, Zhang Y, Pu X, Gao R, Xu Z, Song J. Trends in Herbgenomics. *Sci China*  
596 *Life Sci* 2019; 62: 288–308
- 597 [25] Xu Z, Xin T, Bartels D, Li Y, Gu W, Yao H, et al. Genome analysis of the ancient  
598 tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the  
599 acquisition of desiccation tolerance. *Mol Plant* 2018; 11: 983–94
- 600 [26] Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. *De*  
601 *novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant*  
602 *Cell* 2017; 29: 2336–48.
- 603 [27] Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and  
604 morphinan production. *Science* 2018; 362: 343–7.
- 605 [28] Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the genome sequence  
606 of the medicinal plant *Salvia miltiorrhiza*. *Mol Plant* 2016; 9: 949–52.
- 607 [29] Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, et al. Draft  
608 genome sequence of *mentha longifolia* and development of resources for mint cultivar  
609 improvement. *Mol Plant* 2017; 10: 323–39.
- 610 [30] Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J.  
611 Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin  
612 interactions. *Nat Biotechnol* 2013; 31: 1119–25.
- 613 [31] VanBuren R, Wai CM, Ou S, Pardo J, Bryant D, Jiang N, et al. Extreme haplotype  
614 variation in the desiccation-tolerant clubmoss *Selaginella lepidophylla*. *Nat Commun*  
615 2018; 9: 13.
- 616 [32] Bennetzen JL, Wang H. The contributions of transposable elements to the structure,

- 617 function, and evolution of plant genomes. *Annu Rev Plant Biol* 2014; 65: 505–30.
- 618 [33] Liu C, Cheng YJ, Wang JW, Weigel D. Prominent topologically associated  
619 domains differentiate global chromatin packing in rice from *Arabidopsis*. *Nat Plants*  
620 2017; 3: 742–8.
- 621 [34] Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, et al. Genome-wide  
622 analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res* 2015; 25:  
623 246–56.
- 624 [35] Song C, Liu Y, Song A, Dong G, Zhao H, Sun W, et al. The *Chrysanthemum*  
625 *nankingense* genome provides insights into the evolution and diversification of  
626 chrysanthemum flowers and medicinal traits. *Mol Plant* 2018; 11: 1482–91.
- 627 [36] Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C, et al. Genome sequencing of the  
628 high oil crop sesame provides insight into oil biosynthesis. *Genome Biol* 2014; 15: R39.
- 629 [37] Chiang YC, Huang BH, Liao PC. Diversification, biogeographic pattern, and  
630 demographic history of Taiwanese *Scutellaria* species inferred from nuclear and  
631 chloroplast DNA. *PLoS One* 2012; 7: e50844.
- 632 [38] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The  
633 grapevine genome sequence suggests ancestral hexaploidization in major angiosperm  
634 phyla. *Nature* 2007; 449: 463–7.
- 635 [39] Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the  
636 most recent common ancestor of flowering plants. *Nat Genet* 2017; 49: 490–6.
- 637 [40] Navarro A, Barton NH. Chromosomal speciation and molecular divergence–  
638 accelerated evolution in rearranged chromosomes. *Science* 2003; 300: 321–4.



- 639 [41] Zhang X, Abrahan C, Colquhoun TA, Liu CJ. A proteolytic regulator controlling  
640 chalcone synthase stability and flavonoid biosynthesis in *Arabidopsis*. *Plant Cell* 2017;  
641 29: 1157–74.
- 642 [42] Zhang M, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang HB. Preparation of  
643 megabase-sized DNA from a variety of organisms using the nuclei method for advanced  
644 genomics research. *Nat Protoc* 2012; 7: 467–78.
- 645 [43] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu:  
646 scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat  
647 separation. *Genome Res* 2017; 27: 722–36.
- 648 [44] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:  
649 assessing genome assembly and annotation completeness with single-copy orthologs.  
650 *Bioinformatics* 2015; 31: 3210–2.
- 651 [45] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler  
652 transform. *Bioinformatics* 2009; 25: 1754–60.
- 653 [46] Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an  
654 optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 2015; 16: 259.
- 655 [47] Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-  
656 to-use annotation pipeline designed for emerging model organism genomes. *Genome*  
657 *Res* 2008; 18: 188–96.
- 658 [48] Ghosh S, Chan CK. Analysis of RNA-seq data using TopHat and Cufflinks.  
659 *Methods Mol Biol* 2016; 1374: 339–61.
- 660 [49] Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for

661 comparative genomics. *Gonome Biol* 2019; 20: 238.

662 [50] De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for  
663 the study of gene family evolution. *Bioinformatics* 2006; 22: 1269–71.

664 [51] Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for  
665 detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*  
666 2012; 40: e49.

667 [52] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam  
668 protein families database in 2019. *Nucleic Acids Res* 2019; 47: D427–32.

669 [53] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary  
670 genetics analysis across computing platforms. *Mol Biol Evol* 2018; 35: 1547–9.

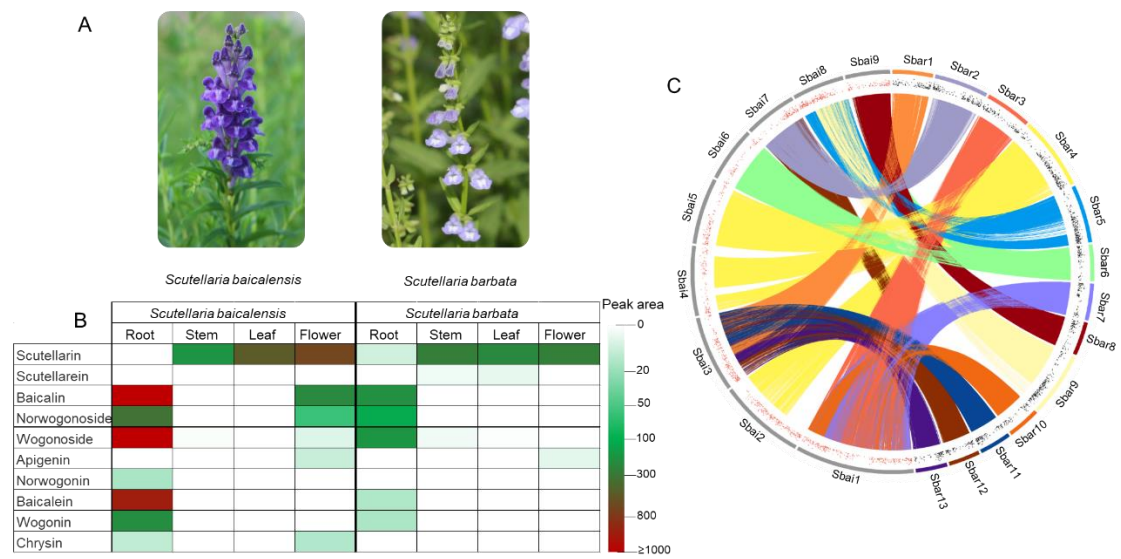
671 [54] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: genome sequence  
672 archive. *Genomics Proteomics Bioinformatics* 2017; 15: 14–8.

673 [55] Zhang Z, Zhao W, Xiao J, Bao Y, Wang F, Hao L, et al. Database Resources of the  
674 BIG Data Center in 2019. *Nucleic Acids Res* 2019; 47: D8–14.

675

676 **Figures**

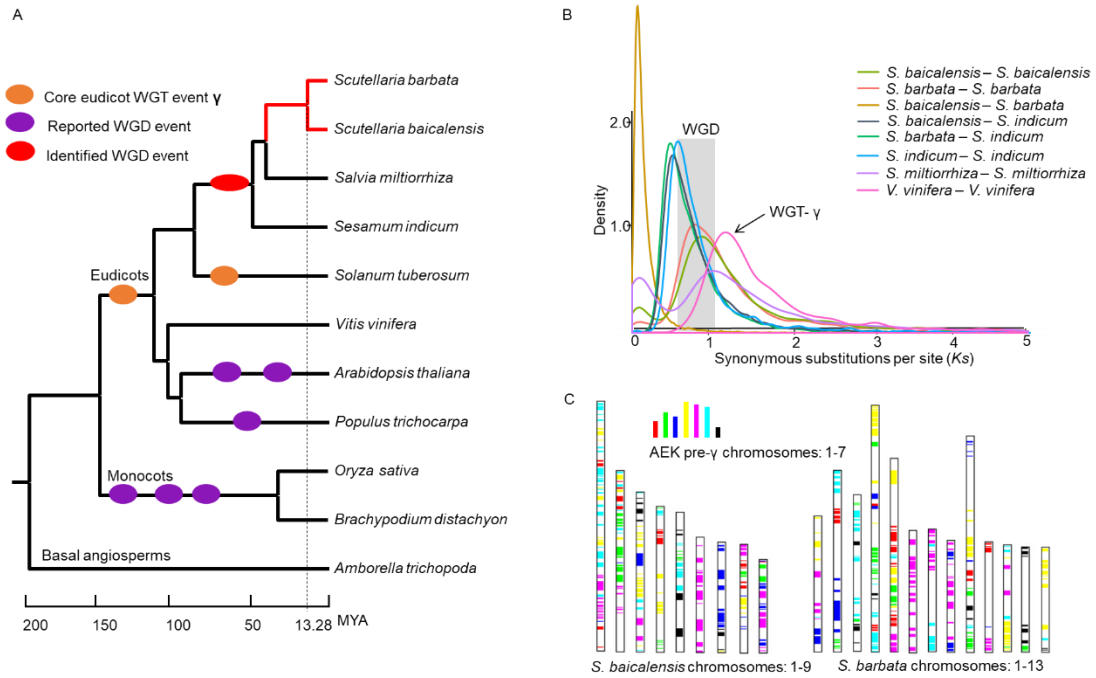
677



678

679 Figure 1. The similar morphology and flavonoid distribution of *S. baicalensis* and *S.*  
 680 *barbata*, as well as their genome collinearity. A. Morphological differences between the  
 681 flowers of *S. baicalensis* and *S. barbata*. B. Content distribution of flavone compounds  
 682 in different tissues of *S. baicalensis* and *S. barbata*, including roots, stems, leaves and  
 683 flowers. C. Comparison of nucleic acid sequences from 9 *S. baicalensis* chromosomes  
 684 and 13 *S. barbata* chromosomes; mapping regions with more than 90% sequence  
 685 similarity over 5 kb were linked. The red and black dots represent significant changes  
 686 in gene expression ( $\text{Log}_2\text{foldchange} > 1$ ,  $\text{FPKM} > 10$ ) in the root tissues of *S. baicalensis*  
 687 and *S. barbata*, respectively.

688

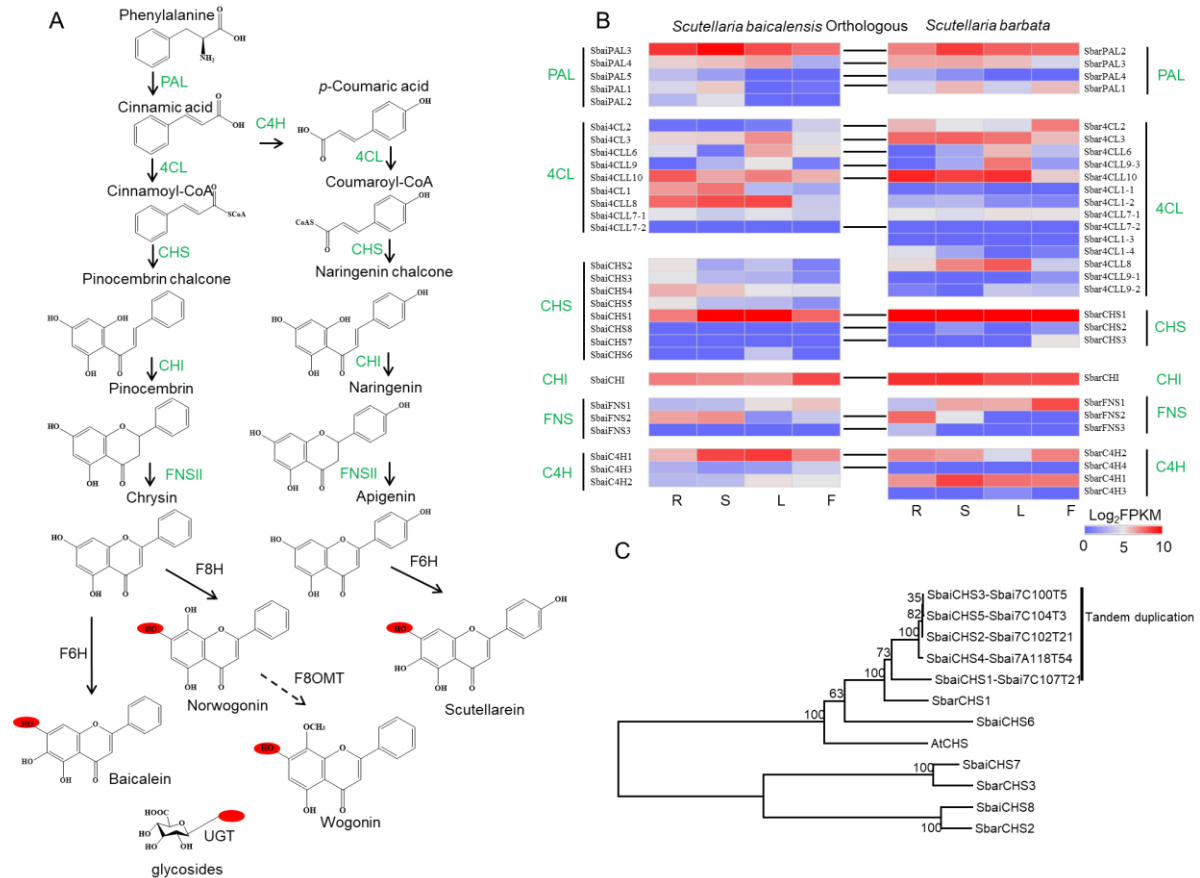


689

690 Figure 2. The divergence time and whole genome duplication of the *S. baicalensis* and  
691 *S. barbata* genomes. A. The phylogenetic tree was constructed using 235 single-copy  
692 orthologous genes from 11 angiosperms. The basal angiosperm *Amborella trichopoda*  
693 was chosen as the outgroup. Speciation times were estimated based on the reported  
694 divergence times *Amborella trichopoda-Vitis vinifera* (173-199 MYA) and *Populus*  
695 *trichocarpa-Arabidopsis thaliana* (98-117 MYA). The orange ovals represented the  
696 reported whole genome triplication events (WGT), and the red and purple ovals  
697 represent whole genome duplication events (WGD). B. Synonymous substitution rate  
698 ( $K_s$ ) distributions of syntenic blocks for the paralogs and orthologs of *S. baicalensis*, *S.*  
699 *barbata*, *S. miltiorrhiza*, *S. indicum*, and *Vitis vinifera*. C. Comparison with ancestral  
700 eudicot karyotype (AEK) chromosomes. The syntenic AEK blocks are painted onto *S.*  
701 *baicalensis* and *S. barbata* chromosomes, respectively.

702

703



704

705 Figure 3. The potential biosynthetic pathway of flavonoids and species-specific gene

706 expansion in *S. baicalensis* and *S. barbata*. A. Biosynthetic genes related to flavones

707 and their glycosides. Phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase

708 (C4H), 4-CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI),

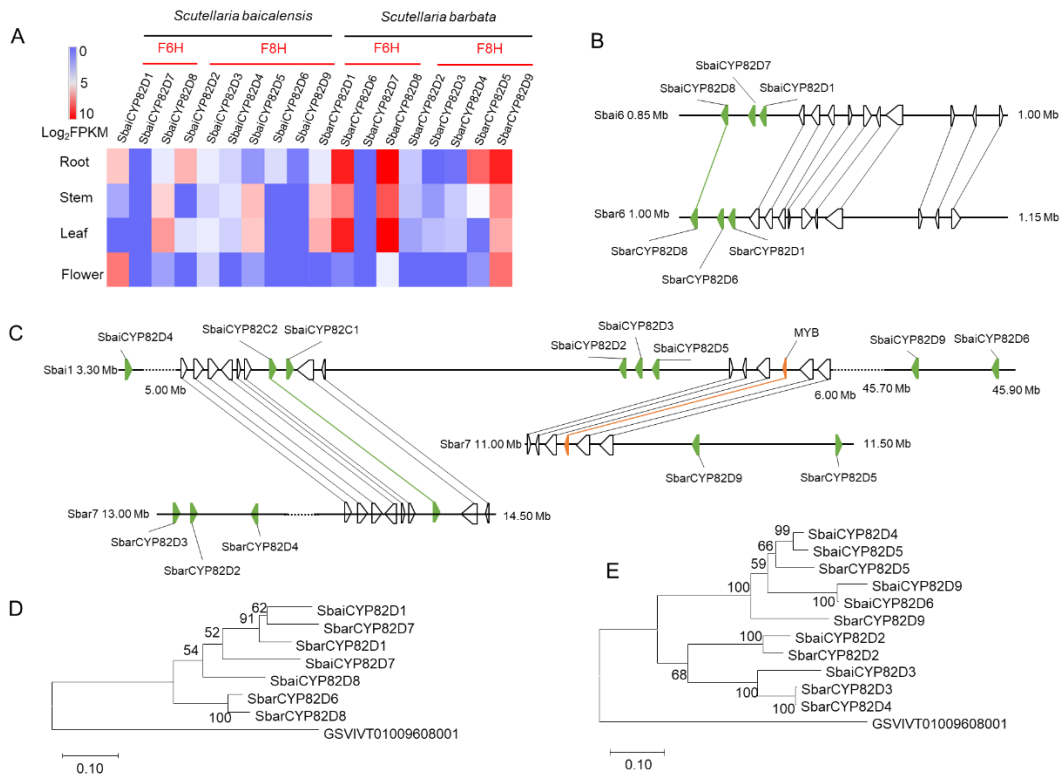
709 flavone synthase (FNSII), flavone 6-hydroxylase (F6H) and flavone 8-hydroxylase

710 (F8H). B. The expression profile and orthologous gene pairs of flavone biosynthetic

711 genes in *S. baicalensis* and *S. barbata*. C. Tandem duplication and phylogenetic analysis

712 of CHS genes.

713



714

715 Figure 4. The tandem repeat of flavone hydroxylase genes in *S. baicalensis* and *S.*  
 716 *barbata* revealed the divergent evolution. A. Identification and expression of CYP82D  
 717 subfamily genes. Flavone 6-hydroxylase (F6H), and flavone 8-hydroxylase (F8H). B.  
 718 Collinearity of CYP82D1 (F6H) regions between *S. baicalensis* and *S. barbata*. C.  
 719 Synteny of CYP82D2 (F8H) regions between *S. baicalensis* and *S. barbata*. D.  
 720 Phylogenetic tree of CYP82D1 groups. The grape CYP82D (GSVIVT01009608001)  
 721 was chosen as outgroup. E. Phylogenetic tree of CYP82D2 groups. The grape CYP82D  
 722 (GSVIVT01009608001) was chosen as outgroup.

723

#### 724 **Supplementary material:**

725 Supplementary Figure S1. Genome size estimation using flow cytometry and the 21 *k*-  
 726 mer distribution. A. Flow cytometry analysis using *Salvia miltiorrhiza* data as internal  
 727 standards. B. The 21 *k*-mer distribution from Illumina short reads of *S. baicalensis* and  
 728 *S. barbata*.

729

730 Supplementary Figure S2. Hi-C intrachromosomal contact map of *S. baicalensis* and *S.*  
 731 *barbata* chromosomes. The red diagonal line indicates a high number of

732 intrachromosomal contacts. A. Hi-C heatmap of *S. baicalensis*. B. Hi-C heatmap of *S.*  
733 *barbata*.

734

735 Supplementary Figure S3. Genome synteny analysis of *S. baicalensis* and *S. barbata*  
736 using MCscanX.

737

738 Supplementary Figure S4. The alignment of large-scale DNA sequences between *S.*  
739 *baicalensis* and *S. barbata* using MUMmer.

740

741 Supplementary Figure S5. Insertion time distribution of intact LTR-RTs in *S.*  
742 *baicalensis* and *S. barbata* assuming a mutation rate of  $\mu=1.3\times 10^{-8}$  (per bp per year).

743

744 Supplementary Figure S6. Genome-wide chromatin packing analysis in *S. baicalensis*.

745 A. The intrachromosomal interactions revealing the A/B compartments of *S.*

746 *baicalensis*. B. The ratio of TE and gene numbers between the A and B compartments.

747 C. The interchromosomal interactions of *S. baicalensis*.

748

749 Supplementary Figure S7. Genome-wide chromatin packing analysis in *S. barbata*. A.

750 The intrachromosomal interactions revealing the A/B compartments of *S. barbata*. B.

751 The ratio of TE and gene numbers between the A and B compartments. C. The

752 interchromosomal interactions of *S. barbata*.

753

754 Supplementary Figure S8. The Gene family expansion and contraction of candidate

755 species. The number of expansion and contraction events of 20 nodes are listed in Table

756 S10.

757

758 Supplementary Figure S9. The grape genome was painted into *S. baicalensis* and *S.*

759 *barbata* genome, respectively. The synteny from paralogs was detected by MCScanX.

760

761 Supplementary Figure S10. The gene syntenic analysis within candidate species. Dot  
762 plot presented that the gene synteny of grape-Sesame, grape-*S. baicalensis*, and grape-  
763 *S. barbata*, respectively. The red circles highlighted the duplication events after WGD-  
764  $\gamma$  event.

765

766 Supplementary Figure S11. Ultraperformance liquid chromatography (UPLC)  
767 detection (280 nm) of flavonoid contents. The UPLC detection of flavonoids in  
768 different tissues of *S. baicalensis* and *S. barbata*, including baicalein, scutellarein,  
769 wogonin, and their glycosides (baicalin, scutellarin, and wogonoside). The compound  
770 information, including detailed retention times and spectrum data, is listed in Table S12.

771 A. Flavonoid contents of *S. baicalensis*. B. Flavonoid contents of *S. barbata*.

772

773 Supplementary Figure S12. Gene expression clusters based on *k*-means in *S.*  
774 *baicalensis*. All expressed genes were clustered into 48 clusters in different *S.*  
775 *baicalensis* tissues, namely, root, stem, leaf, and flower tissues.

776

777 Supplementary Figure S13. Gene expression clusters based on *k*-means in *S. barbata*.  
778 All expressed genes were clustered into 48 clusters in different *S. barbata* tissues,  
779 namely, root, stem, leaf, and flower tissues.

780

781 Supplementary Figure S14. Phylogenetic analysis of PAL, C4H, 4CL, and FNSII from  
782 *S. baicalensis* and *S. barbata* using the maximum likelihood method.

783

784 Supplementary Figure S15. The potential biosynthetic pathway of baicalein,  
785 scutellarein, wogonin, and their glycosides (baicalin, scutellarin, and wogonoside),  
786 catalyzing chrysin and apigenin.

787

788 Supplementary Figure S16. The physical clusters of CYP450s (5 gene clusters per 500  
789 kb) in *S. baicalensis*.

790



791 Supplementary Figure S17. The physical clusters of CYP450s (5 gene clusters per 500  
792 kb) in *S. barbata*.

793

794 Supplementary Figure S18. The phylogenetic analysis of CYP82D, CYP93B, and  
795 CYP73A members from *S. baicalensis* and *S. barbata* using the maximum likelihood  
796 method.

797

798 Supplementary Table S1. The statistics of sequencing data from the SMRT and ONT  
799 platforms and corrected reads using CANU.

800

801 Supplementary Table S2. The assembled statistics of the *S. baicalensis* and *S. barbata*  
802 genome.

803

804 Supplementary Table S3. The genome synteny between *S. baicalensis* and *S. barbata*.

805

806 Supplementary Table S4. Genome annotations among *S. baicalensis*, *S. barbata* and *S.*  
807 *miltiorrhiza*.

808

809 Supplementary Table S5. Annotation of *S. baicalensis* TEs.

810

811 Supplementary Table S6. Annotation of *S. barbata* TEs.

812

813 Supplementary Table S7. Summary of intact LTR retrotransposons in *S. baicalensis* and  
814 *S. barbata*.

815

816 Supplementary Table S8. Annotation of *S. baicalensis* and *S. barbata* rRNA.

817

818 Supplementary Table S9. Identification of SSRs in the *S. baicalensis* and *S. barbata*  
819 genome.

820

821 Supplementary Table S10. Gene family expansion and contraction of candidate species  
822 according to phylogenetic analysis ( $P < 0.01$ ).

823

824 Supplementary Table S11. The  $K_s$  value and divergence time of paralogous or  
825 orthologous gene pairs among *S. baicalensis*, *S. barbata*, *S. miltiorrhiza*, *S. indicum*,  
826 and *V. vinifera*.

827

828 Supplementary Table S12. The compound information of UPLC detection including  
829 retention time and spectrum, is shown in Supplementary Figure S11.

830

831 Supplementary Table S13. The Pfam annotation of genes with high expression in the  
832 roots of *S. baicalensis*.

833

834 Supplementary Table S14. The Pfam annotation of genes with high expression in the  
835 roots of *S. barbata*.

836

837 Supplementary Table S15. The expression of chrysin and apigenin biosynthetic genes  
838 in different organs of *S. baicalensis*.

839

840 Supplementary Table S16. The expression of chrysin and apigenin biosynthetic genes  
841 in different organs of *S. barbata*.

842

843 Supplementary Table S17. The *Ka* and *Ks* analysis of chrysin and apigenin biosynthetic  
844 genes in *S. baicalensis* and *S. barbata*.

845

846 Supplementary Table S18. The expression of CYP82D members in different tissues of  
847 *S. baicalensis* and *S. barbata*.

848

849 Supplementary Table S19. The *Ks* values of gene pairs related to flavone biosynthesis  
850 in *S. baicalensis* and *S. barbata*.

851