Analysis of Genetic Diversity in the Traditional Chinese Medicine Plant ‘Kushen’ (Sophora flavescens Ait.)

Carolyn J. Schultz1*, Shashi N. Goonetilleke1†, Jianping Liang2*, Jelle Lahnstein1†, Kara A. Levin1†, Tina Bianco-Miotto1†, Rachel A. Burton1†, Diane E. Mather1† and Kenneth J. Chalmers1*†

Kushen root, from the woody legume Sophora flavescens, is a traditional Chinese medicine that is a key ingredient in several promising cancer treatments. This activity is attributed in part to two quinolizidine alkaloids (QAs), oxymatrine and matrine, that have a variety of therapeutic activities in vitro. Genetic selection is needed to adapt S. flavescens for cultivation and to improve productivity and product quality. Genetic diversity of S. flavescens was investigated using genotyping-by-sequencing (GBS) on 85 plants grown from seeds collected from 9 provinces of China. DArTSeq provided over 10,000 single nucleotide polymorphism (SNP) markers, 1636 of which were used in phylogenetic analysis to reveal clear regional differences for S. flavescens. One accession from each region was selected for PCR-sequencing to identify gene-specific SNPs in the first two QA pathway genes, lysine decarboxylase (LDC) and copper amine oxidase (CAO). To obtain SfCAO sequence for primer design we used a targeted transcript capture and assembly strategy using publicly available RNA sequencing data. Partial gene sequence analysis of SfCAO revealed two recently duplicated genes, SfCAO1 and SfCAO2, in contrast to the single gene found in the QA-producing legume Lupinus angustifolius. We demonstrate high efficiency converting SNPs to Kompetitive Allele Specific PCR (KASP) markers developing 27 new KASP markers, 17 from DArTSeq data, 7 for SfLDC, and 3 for SfCAO1. To complement this genetic diversity analysis a field trial site has been established in South Australia, providing access to diverse S. flavescens material for morphological, transcriptomic, and QA metabolite analysis. Analysis of dissected flower buds revealed that anthesis occurs before buds fully open suggesting a potential for S. flavescens to be an inbreeding species, however this is not supported by the relatively high level of heterozygosity observed. Two plants from the field trial site were analysed by quantitative real-time PCR and levels of matrine and oxymatrine were assessed in a variety of tissues. We are now in a strong position to select diverse plants for crosses to accelerate the process of genetic selection needed to adapt kushen to cultivation and improve productivity and product quality.

Keywords: quinolizidine alkaloids, traditional Chinese medicine, DART sequencing, lysine decarboxylase, copper amine oxidase, single nucleotide polymorphisms, gene duplication, genetic diversity
INTRODUCTION

Extracts from roots of the woody deciduous legume kushen (Sophora flavescens Ait.) have long been used in traditional Chinese medicine. The medicinal use of kushen root preparations was first documented in 200 A.D. in the Chinese book Shen Nong Ben Cao Jing1 and continues today for treatment of a variety of conditions, including cancers, hepatitis, and heart disease (Sun et al., 2012; He et al., 2015). Of particular importance is the use of kushen extracts in combination with extracts from Smilax glabra in compound kushen injection (CKI). CKI has been used in Chinese hospitals to treat cancer for over 20 years (Cui et al., 2019). When LaLDC was overexpressed in Arabidopsis thaliana (which does not normally make cadaverine), both cadaverine and cadaverine-derived alkaloids were produced (Shimizu et al., 2019). In S. flavescens, an LDC gene (SfLDC) has been isolated and reported to be most abundantly expressed in leaf and stem tissues, with little or no expression detected in inflorescence tissues (pedicels, flower buds, and flowers) or in callus (Han et al., 2015).

Traditionally, kushen roots were harvested from wild populations in China, but with increasing demand for the product, kushen began to be cultivated, especially in the northwestern provinces Shanxi, Hubei, Henan, and Hebei (Weng et al., 2017). Agricultural practices for this new crop remain to be optimised. For example, it has been suggested that seed might be used instead of roots as a source of matrine and oxymatrine (Weng et al., 2017). Further, there may be a need for genetic selection to adapt kushen to cultivation and to improve productivity and product quality. Relatively little is known about the extent of genetic and phenotypic diversity in S. flavescens. In a study of plants from four locations (three in China and one in Taiwan), Lin et al. (2014) established that the plants were diploid (2n = 18) and observed statistically significant variation in biochemical composition (matrine, oxymatrine, flavonoids, and antioxidants) and 11 morphological traits among field-grown plants.

In the research reported here, genetic diversity of kushen was investigated by applying GBS to 85 accessions of S. flavescens obtained from nine provinces of China. Further, to extend understanding of QA synthesis in kushen, a full-length SfCAO coding sequence was isolated from publicly available transcriptome data. A kushen field nursery was established in southern Australia and the expression of SfLDC and SfCAO was investigated in seed-grown plants from that nursery.

MATERIALS AND METHODS

Growth and Collection of Plant Material

Seeds of S. flavescens were obtained from a collection maintained at Shanxi Agricultural University. They were collected from non-cultivated (natural) habitats in nine different provinces in China (counting inner Mongolia as a province), representing the natural distribution of the species. Seed were collected from a single region per province, ranging from Meihekou City, Jilin province in the northeast of China to Anshun City, Guizhou province in the southwest of China (Figure 2A). In each province, seed was collected from multiple plants at a single site and the seed pooled. A subsample of seeds from each population was imported into Australia. A “commercial” sample

1https://en.wikipedia.org/wiki/Shennong_Ben_Cao_Jing (accessed March 12, 2021).

unknown. While transcriptional analysis in lupin has identified numerous genes that may be involved in QA synthesis, pathway regulation or transport, these candidate genes remain to be fully analysed and confirmed (Yang et al., 2017; Kroc et al., 2019).

In L. angustifolius, the genes encoding LDC (LaLDC) and CAO (LaCAO) have been isolated and found to be expressed in leaves and stems (Bunsupa et al., 2012; Yang et al., 2017). When LaLDC was overexpressed in Arabidopsis thaliana (which does not normally make cadaverine), both cadaverine and cadaverine-derived alkaloids were produced (Shimizu et al., 2019). In S. flavescens, an LDC gene (SfLDC) has been isolated and reported to be most abundantly expressed in leaf and stem tissues, with little or no expression detected in inflorescence tissues (pedicels, flower buds, and flowers) or in callus (Han et al., 2015).

Traditionally, kushen roots were harvested from wild populations in China, but with increasing demand for the product, kushen began to be cultivated, especially in the northwestern provinces Shanxi, Hubei, Henan, and Hebei (Weng et al., 2017). Agricultural practices for this new crop remain to be optimised. For example, it has been suggested that seed might be used instead of roots as a source of matrine and oxymatrine (Weng et al., 2017). Further, there may be a need for genetic selection to adapt kushen to cultivation and to improve productivity and product quality. Relatively little is known about the extent of genetic and phenotypic diversity in S. flavescens. In a study of plants from four locations (three in China and one in Taiwan), Lin et al. (2014) established that the plants were diploid (2n = 18) and observed statistically significant variation in biochemical composition (matrine, oxymatrine, flavonoids, and antioxidants) and 11 morphological traits among field-grown plants.

In the research reported here, genetic diversity of kushen was investigated by applying GBS to 85 accessions of S. flavescens obtained from nine provinces of China. Further, to extend understanding of QA synthesis in kushen, a full-length SfCAO coding sequence was isolated from publicly available transcriptome data. A kushen field nursery was established in southern Australia and the expression of SfLDC and SfCAO was investigated in seed-grown plants from that nursery.

MATERIALS AND METHODS

Growth and Collection of Plant Material

Seeds of S. flavescens were obtained from a collection maintained at Shanxi Agricultural University. They were collected from non-cultivated (natural) habitats in nine different provinces in China (counting inner Mongolia as a province), representing the natural distribution of the species. Seed were collected from a single region per province, ranging from Meihekou City, Jilin province in the northeast of China to Anshun City, Guizhou province in the southwest of China (Figure 2A). In each province, seed was collected from multiple plants at a single site and the seed pooled. A subsample of seeds from each population was imported into Australia. A “commercial” sample
Schultz et al. Genetic Diversity in Kushen Plants

FIGURE 1 | Proposed quinolizidine alkaloid biosynthesis pathway. The first two steps in the pathway have been confirmed. The gene for the first step, the conversion of L-lysine to cadaverine by a bifunctional lysine/ornithine decarboxylase (L/ODC), has been cloned from a variety of species including S. flavescens (Bunsupa et al., 2012). The conversion of cadaverine to 5-aminopentanal, requires a copper amino oxidase (CAO) and this intermediate is assumed to be spontaneously cyclised to 1-piperideine. The CAO gene involved in QA biosynthesis has been cloned from L. angustifolius (Yang et al., 2017) but not from S. flavescens. 1-piperideine is the precursor for different QAs in S. flavescens [including (+)-matrine and (+)-oxymatrine] and L. angustifolius (including (–)-lupinine and (+)-epilupinine] (Ohmiya et al., 1995; Sato et al., 2018). Dashed arrows indicate multiple reaction steps.

Identification of Cadaverine/Copper-Containing Amine Oxidase Transcripts From RNAseq Datasets

To obtain a full length cDNA sequence for SfCAO, mirabait transcript capture (Chevreux et al., 2004) was used to identify short reads that were subsequently assembled in Geneious (8.1.9) as follows. Two RNAseq datasets (DRR031281 and DRR031283) were selected from the nine datasets publicly available for S. flavescens (project PRJDB3906) (Han et al., 2015) as they had the highest number of sequences matching the QA pathway genes SfLDC (AB561138.1) and LaCAO (MF152953) (BLASTn, no filtering, word size n = 7, data not shown). The datasets were trimmed and error corrected using Quorum, version 1.1.1 (Marçais et al., 2015) or Trim Galore, version 0.4.2 (Krueger, 2015). Housekeeping genes were also obtained for designing PCR primers using L. angustifolius sequences as bait and the S. flavescens sequence, SfLDC (AB561138.1), was used as a positive control Sophora for the mirabait short-read capture method. The L. angustifolius bait sequences [full coding sequence plus up to 20 base pairs (bp) of the 5′ and 3′ untranslated regions (UTRs) (if available)] were as follows: LaCAO (MF152953), tubulin (XM_019596948.1), cyclophilin (XM_019595391), elongation faction-α (XM_019597475.1), and ubiquitin-60S ribosomal protein (XM_019577360.1). The first round of mirabait (mirabait1) used low stringency parameters {-k 21 [k = k-mer-length (default = 31)]; and –n 2 [n = minimum number of k-mers needed for a sequence to be selected (default = 1)]}. Sequences were assembled in Geneious and contigs evaluated by BLASTn to identify the best sequences. A second round of mirabait was performed (mirabait2) at high

Genomic DNA Extraction for DArT Sequencing

Genomic DNA was extracted from freeze dried young leaf tissue using a modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle, 1987), that was optimised for yield and quality in 2 mL tubes. DNA was sent to Diversity Arrays Technology (Bruce, ACT, Australia) for GBS using the DArTseq platform. Markers scores were provided as single-nucleotide polymorphisms (SNPs) (pairs of sequence tags differing only by 1 nt) or presence/absence polymorphisms detected in some accessions but not others. Two rounds of DArTseq were performed, a preliminary round on four samples and a full round on 89 samples.

Frontiers in Plant Science | www.frontiersin.org 3 August 2021 | Volume 12 | Article 704201

of S. flavescens seeds was also purchased over the Internet. Seeds were sown in 15 cm pots filled with coco-peat and supplied with slow-release fertiliser granules Scotts Osmocote Plus Trace Elements Landscape Formula (Scott’s Australia, Bella Vista, Australia), in a glasshouse under 12 h daylength. Young leaf tissue was harvested from up to 10 seedlings from each region for DNA analysis. Additional seedlings were transferred to a field trial site at the Waite Campus of the University of Adelaide, with five to seven plants per row. An initial planting was done in October 2017, and a second planting was done in October 2018, to replace plants that had died. Plants were drip irrigated daily in summer. Tissues for RNA extraction were harvested on 26 November 2018, except for seed and seedpod tissues, that were harvested on 12 February 2019.

2 Alibaba.com
3 www.diversityarrays.com/technology-and-resources/dartseq/
FIGURE 2 | (A) Collection sites for *S. flavescens* seeds spanning the natural range of the species and representing nine different regions in China. A commercial seed sample was obtained from Internet retailer Alibaba. Regions 4–8 coincide with the silk road (Wang et al., 2019). (B) Phylogenetic analysis of SNP data generated using DArTSeq using a Bayesian inference method in MrBayes v3.2.6. A total of 89 samples were analysed, 85 samples grown from seeds collected from 9 provinces of China (10 accessions from most regions, except region 4, 6, and 8, with 8, 9, and 8 accessions, respectively), and 4 samples from commercial seed. Accessions from each population are labelled in a different colour, and are indicated by region number_sample number (e.g., 1_12). Samples in bold and indicated by an asterisk were used to sequence SfLDC and SfCAO gene fragments. A total of 1636 SNPs was used (those with less than 5% missing data and minor allele frequencies greater than 5%). Values reported on the nodes are Bayesian posterior probabilities and are 1.0 unless otherwise indicated. Scale bar represents the probability of nucleotide substitutions per site. Map image: https://upload.wikimedia.org/wikipedia/commons/d/d2/China_administrative_claimed_included.svg.
stringency (default parameters) using the best contigs from mirabait1, short reads assembled in Geneious, evaluated by BLASTn and the best contigs used for further analysis.

Phylogenetic Analysis
Phylogenetic analysis was performed using the DArTSeq SNP data generated on the 89 DNA samples using MrBayes v3.2.6 (Huelsenbeck and Ronquist, 2001) with the following parameters: a general time reversible model and a gamma shaped distribution of rates across sites function were used. The Markov Chain Monte Carlo (MCMC) was set to two million generations with sample frequency of 100 and 5000 burn-in. The analysis used all SNPs with less than 5% missing data and for which the minor allele frequency was at least 5%. The resulting phylogenetic tree was visualised with FigTree v.1.4.4 (Rambaut, 2018).

Phylogenetic analysis of CAO proteins was performed using a maximum likelihood method with MEGA 10.0.5 (Kumar et al., 2018). Sequences (untrimmed) were first aligned using MUSCLE (Edgar, 2004) using the default settings. The best model (JTT + G) was identified using the find best model option and used to generate a maximum likelihood tree (with 1000 bootstrap replications) using all sites in the alignment.

DNA Sequencing of Genic SfLDC and SfCAO PCR Products
To identify SNPs in the two QA pathway genes, SfLDC and SfCAO PCR products were amplified from genomic DNA using Phusion™ High-Fidelity DNA Polymerase (ThermoFisher, Australia). SfLDC (full length) gene sequences were amplified using SfLDC_F_5pr and SfLDC_R1. Two partial gene fragments were amplified for SfCAO: a 5' fragment using SfCAO_5p_Fd and SfCAO_e1_R, and a 3' fragment SfCAO_e10_F and SfCAO_R1. Primer information is listed in Supplementary Table 1. PCR reactions were prepared (50 μL) using 100 ng of genomic DNA. Thermocycling conditions were: 98°C for 30 s, followed by 30 cycles of 98°C for 10 s, annealing at 65°C for 30 s, with extension at 72°C for 15–25 s per kilobase (kb) of sequence, with a final extension at 72°C for 5 min. For accessions where standard PCR was unsuccessful, presumably due to divergent sequence at one or more primer binding sites, touchdown PCR was used (Korbie and Mattick, 2008) with modified cycling conditions after the initial denaturing. Phase 1 (15 cycles, with annealing temperature starting at 66°C and reducing 0.5°C per cycle), followed by Phase 2 (20 cycles, with annealing at 59°C).

For Sanger sequencing, PCR products were purified using the NucleoSpin gel kit from Macherey-Nagel (Bethlehem, PA, United States), or by gel isolation when two or more PCR products were produced (see section “Results”), as per manufacturer’s instructions. PCR products were sequenced using both forward and reverse primers. Sequencing chromatograms were analysed using Geneious 8.1.9. The following additional primers were used to sequence the genomic fragments: SfLDC_F314, SfLDC_F752, SfLDC_R596, SfLDC_R1026 and in some cases SfCAO_e11_F and SfCAO_e11_R (Supplementary Table 1).

Development of KASP Markers
Kompetitive Allele Specific PCR (KASP) uniplex assays (hereafter KASP markers) (He et al., 2014) were developed and tested using the sequences derived from 19 randomly selected DArTseq-derived GBS tags and two QA pathway genes, SfLDC and SfCAO1. Primer sets were designed using Kraken™ software (LGC Limited, Teddington, United Kingdom). The KASP markers for GBS-derived SNPs were assayed on the four DNA samples that were used in the preliminary round of GBS analysis. The KASP markers for SNPs in SfLDC and SfCAO1 were assayed on 86 of the DNA samples used in the second round of GBS analysis. KASP genotyping was conducted using an automated SNPLine system (LGC Limited, Teddington, United Kingdom).

RNA Extraction, cDNA Synthesis, and Quantitative Real-Time PCR
Tissues for RNA were harvested, immediately frozen in liquid nitrogen and stored at −80°C until required. Tissues (leaf, stem, flower buds, and open flowers and roots) were sampled from two field grown plants, 1_LC1 [Lingyuan City (LC), Liaoning Province] and 9_AC1 [Anshun City (AC), Guizhou Province] in triplicate, without destroying the plants (Supplementary Figures 1, 2). Seedpods containing seeds were collected from another plant, 1_LC4 (Supplementary Figure 3). Frozen tissue was ground to a fine powder in the presence of liquid nitrogen, either in a mortar and pestle or an IKA A11 grinder [IKA® Works (Asia) Sdn Bhd, Selangor, Malaysia]. Total RNA was extracted from 100 mg of frozen ground tissue using the Spectrum™ Plant Total RNA Kit according to manufacturer’s instructions. Total RNA was treated with DNase I (New England Biolabs, M0303S, Ipswich, MA, United States), then 750 ng of RNA was used to prepare cDNA using Superscript™ III reverse transcriptase (ThermoFisher) using oligodT18 as primer (50°C for 1 h).

Transcript profiling of SfLDC and SfCAO was carried out with all cDNA samples using quantitative real-time PCR (qPCR) methods as described (Burton et al., 2011), and gene-specific primers listed in Supplementary Table 1. Data was normalised against the geometric mean of the four control genes: cyclophilin (CYC), tubulin (TUB), elongation factor-α (EFa), and ubiquitin-60S ribosomal protein (UBQ) (Burton et al., 2011).

Full length SfCAO cDNA sequences were obtained by PCR from one of three stem cDNA samples each for 1_LC and 9_AC. PCR was performed with Phusion as described above except that a longer extension (60 s) was used with primers SfCAO_5p_Fd and SfCAO_R1. PCR products were sequenced (Sanger) with the PCR primers and additional internal primers: SfCAO_e5_F, SfCAO_ex8_F, SfCAO_ex7_R, and SfCAO_ex4_R (Supplementary Table 1).

Extraction and Quantification of Oxymatrine and Matrine
Protocols for extraction of oxymatrine and matrine from S. flavescens tissues and a reversed-phase high performance liquid chromatography (HPLC) quantification method were based on the Hong Kong Chinese Materia Medica Standards (HKCMMS) (HKCMMS, 2012), with modifications for 50 ± 5 mg tissue (exact

---

AGRF.com.au

---

Oxymatrine and Matrine
ammonium hydroxide solution (28%, v/v) as before, and the centrifugation (13,000 \times 30 \text{ min}) in a 65W sonication bath (at room temperature). After vortexing (10 s) to resuspend the pellet followed by 5 min in a 65W sonication bath (at room temperature). After centrifugation (20,000 \times 5 \text{ min}) and 100 \mu L of supernatant was removed and fully dried in a SpeediVac (Savant). The remaining supernatant was removed and stored at –20°C, and subsequently used to prepare concentrated extracts from some samples by drying down 1 mL of soluble extract. Dried samples were resuspended in 30–50 \mu L 30% ethanol (v/v) and clarified by centrifugation (20,000 \times 5 \text{ min}) and 100 \mu L of supernatant was removed and fully dried in a SpeediVac (Savant). The remaining supernatant was removed and stored at –20°C, and subsequently used to prepare concentrated extracts from some samples by drying down 1 mL of soluble extract. Dried samples were resuspended in 30–50 \mu L 30% ethanol (v/v) and clarified by centrifugation (20,000 \times 5 \text{ min}) and 100 \mu L of supernatant was removed and fully dried in a SpeediVac (Savant). The remaining supernatant was removed and stored at –20°C, and subsequently used to prepare concentrated extracts from some samples by drying down 1 mL of soluble extract. Dried samples were resuspended in 30–50 \mu L 30% ethanol (v/v) and clarified by centrifugation (20,000 \times 5 \text{ min}) for transferring to HPLC vials for use and/or short-term storage at –20°C. QAs were extracted from additional aliquots of tissues used for qPCR and two control samples [dried kushen root purchased from a traditional Chinese medicine store in Adelaide (September 2019) and dry seeds, purchased online from footnote 1].

Quinolizidine alkaloids were analysed using an Agilent 1200 HPLC unit, with a diode array detector (set to record at 220 nm) and a Zorbax Eclipse Plus C18 rapid resolution column (HD 2.1 \times 50 \text{ mm}, 1.8 \mu m, Agilent, CA, United States). HPLC was performed using a mobile phase composed of A (0.3% phosphoric acid and 0.912% triethylamine, pH 6.0 in milliQ water) and B (0.3% phosphoric acid and 0.912% triethylamine, pH 6.0 in 70% acetonitrile) with a gradient elution from 4% A to 10% B over 10 min, flow rate 0.4 mL/min, column temperature 35°C. The injection volume was 2 \mu L. Sigma standards were used to establish identity of oxymatrine (O0891) and matrine (M5319) based on their relative retention times and published UV spectra (Ling et al., 2007). A standard curve was produced from a mixture of oxymatrine and matrine, with oxymatrine concentrations of 50, 100, 200, 400, and 800 ng/\mu L and matrine of 2.5, 5, 10, 20, and 40 ng/\mu L in 30% ethanol (from low to high, respectively), reflecting the higher levels of oxymatrine to matrine expected in root tissues.

RESULTS

**DArT Sequencing of Nine Populations of *S. flavescens* From China and Genetic Diversity Analysis**

An initial round of DArTseq using four DNA samples (one from each of Liaoning, Jilin, Shaanxi, and Guizhou provinces: regions 1, 3, 7, and 9, respectively, Figure 2A) yielded 2212 SNP-bearing tag pairs, 1392 of which were scored for all plants. Of the genotypes scored for these 1392 SNPs, 78.7% were homozygous and 21.3% were heterozygous. Nineteen SNPs were selected for which all four plants exhibited homozygosity (two of each genotype or three of one genotype and two of the other). KASP primer sets (He et al., 2014) were designed for these SNPs and applied to the same four DNA samples. Fifteen of these markers clearly distinguished the two alternative homozygous genotypes (Supplementary Table 2). Of the 60 possible genotypes (15 markers \times 4 plants), 58 were scored based on the KASP markers (exceptions: Sf17 and Sf19 for the plant from region 1). Of these, 52 (90%) were identical to the homozygous genotypes from the DArTseq analysis, five (Sf5, Sf7, Sf10, and Sf15 for the plant from region 1; Sf17 for the plant from region 3) were heterozygous and only one (Sf6 for the plant from region 9) contradicted the result from the DArTseq analysis (A:A vs. T:T). Failure of the DArTseq analysis to detect heterozygosity may be due to sequencing depth differing between alleles, with one allele failing to meet the threshold required to call a heterozygote.

DArTseq analysis performed with 89 DNA samples produced 10,599 SNPs, of which 1636 had less than 5% missing data and minor allele frequencies greater than 5%. With phylogenetic analysis of the data for the 1636 SNPs, most plants from each province grouped together and plants from neighbouring provinces clustered together (Figure 2B). There were, however, some exceptions to the tight clustering of samples from the same region. For example, the 10 plants from region 5 (Huaizhou County in Beijing province) were distributed across three clades: one of six plants, another of three plants, and there was a single plant that grouped with most of the accessions from region 6 (Qin County in Shanxi province). These exceptions could be due to human movement of plants or seeds and/or selection of specific characteristics. Overall, it seems that accessions from Qin County (Shanxi province, region 6) may have diverged early and that different changes occurred in the north–east (regions 1–4) and south–west (regions 7–9) of China.

It was observed that anthesis can occur before flowers open (Figure 3), indicating that *S. flavescens* could be capable of self-pollination. Among 1636 SNPs with less than 5% missing data and with minor allele frequencies greater than 0.05 (Supplementary Table 3), the overall frequency of heterozygosity was 0.21. Among these SNPs, 1070 (65.4%) had genotypic frequencies that deviated significantly (p < 0.05) from Hardy–Weinberg expectations, according to Chi-square tests conducted with one degree of freedom. For 1053 of these markers (98.4%), the observed frequency of heterozygotes was lower than the expected frequency (Figure 4).

**Obtaining Full Length Sequence for SfCAO**

To develop gene-specific SNP markers for *LDC* and *CAO* we needed to sequence both genes, however, there was no full-length sequence for *SfCAO* in the NCBI databases. We used mirabait (Chevreux et al., 2004) at low stringency to identify short read RNA sequences matching *SfCAO* from two publicly available but not assembled datasets for *S. flavescens* (Han et al., 2015). After assembly and analysis, the best contig was used in a second round of mirabait to obtain a full length *SfCAO* cDNA sequence from dataset DRR031281 (Supplementary Figure 4A). It was 86.9% identical to the coding sequence of *LaCAO* (MF152953) and consisted of 104 bp of 5' UTR, 2343 bp of coding sequence,
and 169 bp of 3’ UTR (Supplementary Figure 4A). Alignment of the encoded protein with LaCAO and a CAO protein from Cicer arietinum (chickpea) shows that the protein encoded by the SfCAO cDNA contains the important catalytic residues and a C-terminal peroxisomal targeting signal as expected (Reumann et al., 2012; Figure 5A). Phylogenetic analysis provides strong support that SfCAO is the orthologue of LaCAO involved in QA biosynthesis (Yang et al., 2017; Figure 5B).

Full length transcripts were also obtained by mirabait for SfLDC and different variants (1 and 2) were obtained from the two datasets, DRR031281 and DRR031283, respectively. The variants were greater than 98% identical to the SfLDC coding sequence at NCBI (AB561138.1) and differ from each other only at 4 positions (data not shown).

SNPs Identified for the SfLDC Gene

Ten DNA samples were selected for SNP discovery in SfLDC, one from each regional clade and one commercial sample (Figure 2B). For SfLDC, in which no introns were expected, PCR primers were designed in the 5’ and 3’ UTRs of SfLDC to enable sequencing of the full length SfLDC gene (SfLDC_F_5pr and SfLDC_R1, respectively, Supplementary Table 1). The forward primer SfLDC_F_5pr was designed based on sequence alignment between SfLDC AB561138.1, several short reads from S. flavescens...
dataset DRR031281 and a 5' UTR sequence from *S. alopecuroides LDC* (KY038928.1) (Supplementary Figure 5). The reverse primer *SfLDC_R1* was designed based on *SfLDC* AB561138.1. For eight of 10 samples, PCR products of the expected size were obtained with standard PCR cycling conditions (annealing at 65°C). For plants 4_15 and 9_8, standard cycling at a reduced annealing temperature of 55°C resulted in multiple bands (not shown) whereas touchdown PCR (Korbie and Mattick, 2008) provided a single amplification product of the expected size. There may be one or more mismatches between the primer(s) and the template DNA for these two samples.

The SNPs identified in *SfLDC* by Sanger sequencing are detailed in Supplementary Table 4, with example chromatograms in Supplementary Figure 6A. *SfLDC* genomic DNA sequences of the 10 samples are reported in Supplementary Figure 4B and aligned in Supplementary Figure 7A. A total of 14 SNPs were identified, of which six are predicted to result in amino acid changes in the encoded protein (Supplementary Figure 7B). There were several highly polymorphic SNPs where both homozygotes and the heterozygote were observed (positions 41, 381, 459, and 1341, Supplementary Figure 7A).

**PCR Identifies Two Recently Duplicated Genes, *SfCAO1* and *SfCAO2***

The second gene in the QA pathway, CAO, is expected to have 12 exons (11 introns), based on the 6520 bp lupin CAO gene sequence (Lup000530), with one very long intron (intron 1, 1931 bp) (Supplementary Figure 8). For this gene, PCR primers were designed to amplify a short partial 5’ sequence (5’ UTR and most of exon 1) and a 3’ partial sequence (part of exon 10 through to the 3’ UTR).

A 5’ product of the expected size (247 bp) was obtained and sequenced, revealing two polymorphic positions (Supplementary Figures 6B,C and Supplementary Table 5). Surprisingly, all samples were heterozygous at the first SNP (position 81) and 9 of 10 samples were heterozygous at the second polymorphic position (position 220). A plausible explanation for the high heterozygosity became apparent when the 3’ partial fragment was amplified, as two strong products were produced in most samples (Supplementary Figure 9A), indicating that there may be two closely related CAO genes, hereafter called *SfCAO1* and *SfCAO2*, that share the same primer binding sites. Most samples produced a lower band of the expected size ≈ 720 bp (compared to the lupin gene Lup000530, ≈750 bp) and all samples produced an upper band (≈1200 bp).

Sample 4_15 did not produce the 720 bp band and sample 9_8 produced a smaller PCR product, ≈550 bp) (Supplementary Figure 9A). Sequencing revealed that the difference between the 1200 and 720 bp PCR products is due to longer introns in *SfCAO2* compared to *SfCAO1*. For example, in sample 1_12, *SfCAO2* has the same exon sequence as *SfCAO1*, however, intron 10 is 40 bp longer (at 281 bp) than for *SfCAO1* and intron 11 is 399 bp longer (at 546 bp) (Supplementary Figures 9B–D).

Sequencing of the smaller than expected PCR product for *SfCAO1* in sample 9_8 (Supplementary Figure 9A) showed that the reduced size is due to a 181 bp deletion (from position 271 in intron 10 to position 451 in intron 11) that results in loss of exon 11 (Supplementary Figure 7D). Exon 11 is expected to be critical for CAO activity as it contains the codon for one of three catalytic His residues in the CAO domain (Pilam PP01179) (Supplementary Figure 8B). A PCR check showed that seven of nine other plants from region 9 also have the 181 bp deletion (Figures 6A–C). The two samples that do not have the deletion (that is they contain the full length product including exon 11 sequence) are sample 9_10, the more ancestral sample of the main group from region 9, and an “outlier” sample, sample 9_6, which grouped with 6_13 in the phylogenetic tree produced with 1636 DArTSeq SNPs (Figures 2B, 6D,E).

Two of six samples tested from region 4 did not produce a PCR product for the 3’ fragment of *SfCAO1* with the primers used (Figure 6B). These two samples, 4_15 and 4_13, are more closely related to each other than the other four samples tested from region 4 based on phylogenetic analysis (Figures 2B, 6D,E).

**SNPs Identified for *SfCAO1* and *SfCAO2***

Single nucleotide polymorphism were identified in both *SfCAO1* and *SfCAO2* (Supplementary Figures 4D,E, 7D,E and
**FIGURE 5** | Sequence and analysis of deduced SfCAO protein sequence derived from the targeted capture and assembly of *S. flavescens* short read sequences matching LaCAO. (A) SfCAO was aligned with LaCAO (XP_019417507.1) and Cicar_XP_004501882.1 using Muscle (https://www.ebi.ac.uk/Tools/msa/muscle/) and the resulting alignment formatted in Geneious. The box indicates the conserved NYE/X motif and the three histidines (*) indicate the residues that interact with the catalytic copper ion (Yang et al., 2017). The horizontal line at the C-terminus indicates a predicted peroxisomal targeting signal peptide (Reumann et al., 2012).

(B) Maximum likelihood tree (MEGA) of CAO genes from *S. flavescens* (Sf, green), *L. angustifolius* (La, blue), *C. arietinum* (Cicar, chickpea, brown), *Glycine max* (Glyma, black), and *Arabidopsis thaliana* (At, orange). Species known to produce QAs (Sf and La) are in bold. Nodes with ≥60% support (1000 bootstrap replicates) are indicated. The three sequences that are labelled LaCAO protein represent the CAO protein involved in QA biosynthesis from three different sources: ATY35659 [protein encoded by the cDNA sequence reported in Yang et al. (2017)], XP_019417507.1 (NCBI), and the protein predicted from the gene, Lup000530 as used to predict intron positions (legumeinfo.org, with some manual reannotation, see Supplementary Figure 8). One of 10 Arabidopsis thaliana CAO proteins (NP_181777.2 (= AtCuAO3), encoded by At2g42490) was selected based on data showing that the protein groups with other legume CAOs and the gene encoding this protein is the only *Arabidopsis* CAO gene that has 11 introns as found in lupin (Tavladoraki et al., 2016). The scale represents the number of amino acid substitutions per site.
Supplementary Tables 6, 7). Both genes contained a simple sequence repeat (SSR) polymorphism (CT)n near the start of intron 10, for which the number of CT repeats ranged from 6 to 8 in SfCAO1 and from 7 to 9 in SfCAO2 (Supplementary Figure 10 and Supplementary Table 8). Samples that were heterozygous for repeat length could not be fully sequenced due to the frameshift created (Supplementary Figures 7D,E, 10). More SNPs were identified in SfCAO1 than in SfCAO2, including three in the coding sequence of SfCAO1 (Supplementary Tables 6C,D). One of these changes, [G/A]AG, at position 335 results in a non-conservative change from Glu to Lys in the encoded protein.

Single nucleotide polymorphisms were also identified between two full length cDNAs for SfCAO. Prior to our discovery of two putative SfCAO genes involved in QA metabolism, we had collected tissue from two field grown plants, 1_LC1 and 9_AC1, for expression analysis (qPCR) and quantification of secondary metabolites (see section “Tissue-Specific Expression of Two QA Pathway Genes and Accumulation of Oxymatrine”). One of these plants, 9_AC1 was grown from seeds collected in region 9, and therefore potentially contained the deletion variant of SfCAO1 missing exon 11 (Figure 6A), so may not produce a full-length cDNA. Genomic DNA was extracted from 9_AC1 and PCR showed that this sample has the deletion variant of SfCAO1 (Supplementary Figure 11), however, both 1_LC1 and 9_AC1, produced full length SfCAO cDNA sequences. In both cases, only a single PCR product of the expected full-length size was produced (≈2350 bp, Supplementary Figure 11). Sequencing confirmed a full-length CAO cDNA in stem tissue from sample 9.AC1 (Supplementary Figure 4F), indicating that SfCAO2 is expressed. Absence of a smaller cDNA PCR product in 9.AC1 indicates that the deletion variant of SfCAO1 is not expressed at detectable levels in stem tissue. A total of five SNPs were identified between 1.LC1 and 9.AC1 cDNA sequences and a further five were identified compared to the SfCAO sequence assembled from the RNASeq data (Supplementary Table 9).

Gene-Specific KASP Markers for SfLDC and SfCAO
To confirm the utility of the SfLDC and SfCAO SNP data we designed gene specific KASP markers. Seven SfLDC-specific markers and three SfCAO1-specific markers were successfully designed and used to genotype 87 of the 89 samples analysed by DArTseq (Supplementary Table 3). All of the relevant polymorphisms identified by Sanger sequencing (Supplementary Tables 4, 6) were confirmed by the KASP markers, bringing the total number of KASP markers developed to 27 (Supplementary Tables 2, 10).

Tissue-Specific Expression of Two QA Pathway Genes and Accumulation of Oxymatrine
To determine where in S. flavescens plants SfLDC and SfCAO are expressed, a tissue series of cDNA samples was developed from 2660 DArTseq SNPs. (E) A scaled down copy of the entire tree from Figure 2B (to provide context).
whereas SfCAO, but not SfLDC, was highly expressed in stems of 9_AC1, at 7.3 ± 1.1 × 10^3 units (Figure 7). Surprisingly, the expression level of SfLDC in stems of 9_AC1 plants was over 500-fold lower at 11 ± 2 units. Expression of the genes in the other tissues was generally low (<100 units), with the exception of expression of SfCAO in leaves of both 1_LC1 and 9_AC1 (at 544 ± 138 units and 125 ± 26 units, respectively), and root (small) 9_AC1 (110 ± 28 units).

Quinolizidine alkaloid metabolites were extracted, and their levels were quantified to determine if oxymatrine and matrine were produced in young plants growing in Australia. Samples used were subsamples of the tissues used for qPCR. Oxymatrine was detected at low levels in samples from both large and small roots (ranging from 0.11 to 0.18% fresh weight) and at higher levels in immature seeds (0.32% fresh weight) (Figure 8). Matrine was only detected in one of the field samples, the immature seeds (0.007 ± 0%), at similar levels to that found in the kushen root (dried) purchased from a TCM store (0.004 ± 0%). The highest level of matrine was found in the commercial sample of dry S. flavescens seeds (0.26 ± 0.02%) (Supplementary Figure 12).

DISCUSSION

Our research provides much needed information on the molecular diversity of S. flavescens plants from a large part of its observed range within China. A combination of techniques has provided resources for the future, with over 10,000 SNPs identified and 27 of these validated through the development of high throughput KASP markers. These and the other results are discussed below in the context of the existing literature and future plans to achieve more control of secondary metabolite production in S. flavescens through the selection of diverse and genetically defined germplasm.

New information was provided for SfCAO, the second gene in the QA pathway, with its cDNA sequence obtained using targeted capture and assembly from publicly available RNASeq data using the lupin LaCAO sequence (Han et al., 2015; Figure 5). LaCAO was recently identified as one of several genes involved in the domestication of sweet lupin varieties, with low LaCAO gene expression contributing to the sweet/low alkaloid phenotype (Plewiński et al., 2019). We have identified a key difference between lupin and S. flavescens with the duplication of the CAO gene. We found no evidence for another lupin gene (data not shown) that groups with the LaCAO gene QA pathway (Figure 5B) in the available lupin genomes [at NCBI, Legume Information System and the recently published lupin genome assembly (Wang et al., 2021)].

Gene duplication is a key process in the evolution and diversification of secondary metabolite production by providing opportunities for either sub-functionalization (greater substrate specificity or more restricted range of substrates) and/or neo-functionalisation (new activities) (Michael, 2017). Most of the plants from region 9 (Anshun City, Guizhou Province) have a 181 bp deletion in SfCAO1 that results in the loss of exon 11 sequence (and part of the flanking intron sequences) (Figure 6A) which encodes one of three catalytic His residues (Figure 3 and Supplementary Figure 8B). This deletion would almost certainly abolish CAO activity, possibly making this change an evolutionary “dead-end.” We did not detect expression of the deletion variant of SfCAO1 but we found that a full-length cDNA, presumably from SfCAO2, was expressed in stem tissue of sample 9_AC1 (Supplementary Figure 11).
sequences for \textit{SfCAO1} have obvious advantages including obtaining full length genomics isoform sequencing (Hardwick et al., 2019), genome sequencing (Wang et al., 2021), and/or \textit{SfCAO1} presence or absence of primers can be designed and tested and if that is not successful, \textit{SfCAO2} from 1\textsubscript{LC1} (\textit{SfLDC}) or one of the duplicated genes. Although touchdown PCR worked to amplify \textit{SfLDC} from 4\textsubscript{15}, it was not successful at amplifying the \textit{SfCAO1} from 4\textsubscript{15} (Supplementary Figure 9A) or \textit{SfCAO2} from 1\textsubscript{LC1} (Supplementary Figure 11A). New \textit{SfCAO} primers can be designed and tested and if that is not successful, other techniques would have to be explored to determine the presence or absence of \textit{SfCAO1} such as Southern blot analysis (Southern, 2000), genome sequencing (Wang et al., 2021), and/or isoform sequencing (Hardwick et al., 2019).

Full genome sequencing of a variety of accessions would have obvious advantages including obtaining full length genomics sequences for \textit{SfCAO1} and \textit{SfCAO2} (expected size \(\approx 6500\) bp). The coding sequences obtained in this work, part of exons 1 and 10 and all of exons 11 and 12, were identical between the two genes in most varieties, noting that in some samples not all of exon 11 could be sequenced due to frameshifts created by SSRs in the introns (Supplementary Figures 4D,E, 7D,E, 10). Full length genomic sequence would also allow us to determine: (1) whether the apparent SNPs in exon 1 (Supplementary Figures 6B,C) are actually genic differences in \textit{SfCAO1} and \textit{SfCAO2} and (2) whether there are sequence differences between the two full length coding sequences that could lead to functional or regulatory differences between the encoded proteins. The second question in particular could have important implications for QA metabolite production.

The utility of the DARtSeq GBS approach was clearly demonstrated with the finding of over SNPs (1636 SNPs with <5% missing data and minor allele frequencies > 5%) providing robust data for phylogenetic analysis (Figure 2B). By selecting one sample from the main cluster of each region for more detailed sequencing of \textit{SfLDC} and \textit{SfCAO} genic sequences we were able to identify 14 SNPs in \textit{SfLDC} and even more (16) in \textit{SfCAO1} by sequencing just 4 of the 11 exons. Fewer SNPs (8) were identified in the corresponding region of \textit{SfCAO2} despite its longer introns.

High heterozygosity is one of many recognised challenges facing plant breeding of medicinal plants (Wang et al., 2020), and one expected for \textit{S. flavescens} based on anecdotal evidence that it was an outbreeding species. However, in attempting to make crosses, we noted that anthesis can occur before flowers open (Figure 3), raising the possibility that the species might be self-pollinated. Analysis of SNP genotypic data for 85 plants from nine regions in China revealed many significant deviations from Hardy–Weinberg equilibrium and confirmed that heterozygosity was often lower than expected based on allele frequencies. However, the overall frequency of heterozygosity (0.21) was much higher than would be expected with complete or nearly complete self-pollination and the deviations from Hardy–Weinberg equilibrium are likely to be mainly due to population structure, with allele frequencies varying among regions due to selection or genetic drift, in combination with some degree of reproductive isolation. Nevertheless, it would be interesting to investigate whether self-pollination occurs and whether it leads to inbreeding depression. To determine if self-pollination occurs would require knowledge of whether the stigma is receptive before and/or after the flowers fully open. This information could help determine which breeding methods could be implemented for genetic improvement of \textit{S. flavescens}.

A reasonable number of the SNPs detected in coding sequences result in amino acid changes [6 of 13 from \textit{SfLDC} gDNA (Supplementary Table 4) and 3 of 5 from \textit{SfCAO1} gDNA (Supplementary Table 5) and 6 of 15 from \textit{SfCAO} cDNA sequences (Supplementary Table 9)]. Not all the amino acid changes are conservative and therefore any future work should consider molecular modelling to determine the impact of these changes on enzyme function.

To the best of our knowledge, this is the first published report of qPCR based expression data for QA pathway genes in \textit{S. flavescens}. Semi-quantitative RT-PCR was performed by Han et al. (2015) for \textit{SfLDC} comparing three tissues (leaf, stem,
and root) and they observed about twofold higher SfLDC levels in leaves compared to stems, which was consistent with their RNAseq data. In contrast, we observed higher levels in stems than in leaves. For a plant from region 1 (1_LC1) we observed over $2 \times 10^5$ normalised transcript levels for both SfLDC and SfCAO in stem tissue, whereas for the sample from region 9 (9_AC1) there were high levels of SfCAO expression but almost non-detectable levels of SfLDC expression. With the exception of the low SfLDC transcript expression in the stems of 9_AC1 all of the other data is as expected and consistent with the expectation that LDC and CAO are expressed in green tissues, predominantly leaves, and stems with generally lower expression in seedpods, with very low levels in roots, flowers, and seeds (Han et al., 2015; Yang et al., 2017; Otterbach et al., 2019).

The tissue series we developed will provide a quick way to evaluate other candidate genes involved in control of QA metabolism. Co-expression studies with LDC/CAO in lupin (and LDC in S. flavescens) have identified a growing number of candidate genes involved in QA accumulation in as yet unknown ways. The candidate genes include transcription factors, transporters, and a range of potential genes likely involved in metabolite modification (Han et al., 2015; Yang et al., 2017; Kroc et al., 2019; Plewiński et al., 2019; Kamphuis et al., 2020). For candidate genes identified in lupin, we have demonstrated that a mirabad transcript capture (Chevreux et al., 2004) and assembly strategy is an efficient way to use existing resources to obtain transcripts for candidate genes.

For kushen products to be acceptable for use as a TCM, it is required by the Chinese Pharmacopoeia Commission to contain a 1.2% total level of matrine and oxymatrine [reviewed in He et al. (2015)]. The kushen root (TCM-root) we purchased from a reputable TCM store in Australia had barely detectable levels of matrine (0.004%), and the measured oxymatrine level of $0.36 \pm 0.01\%$ is about threefold lower than the expected minimum level, highlighting the need for a better understanding of the genetic and environmental factors controlling QA metabolism. The relatively young field grown plants ($\approx 1.5$ years old) had even lower levels of oxymatrine (and no matrine) (Figure 8). It is widely believed that young S. flavescens plants do not produce as much matrine and oxymatrine as older plants, although we could not find any English scientific literature that documents the actual timing of increased production. A recent paper on the flavonoids produced by S. flavescens showed statistically significant differences from year 1 to year 6 (Chen et al., 2020).

We are confident that if matrine was present in the Australian-grown root samples, it would have been detected as we could detect it in the commercial dried seeds and the immature seeds from plant 1_LC4. We had no issues with the stability or oxidation of the matrine control (Sigma) even after multiple freeze thaw cycles ($-20^\circ C$) and we had good recovery of matrine in a mock extraction control with the QA extraction and storage protocols (data not shown). Our preliminary results showed little difference in oxymatrine concentration between big tuberous roots and small roots (Figure 8). If this can be confirmed with more samples, it will make future phenotyping for QAs more efficient and less disruptive for the plant as it will be possible to harvest smaller (lateral) roots rather than the initial strategy used here of sawing the large root mass in half (Supplementary Figure 1).

We are now well positioned to perform more detailed phenotyping of field grown plants in future research. While evaluation of plant growth, metabolite profiling and transcript analysis can be time consuming and expensive, they have the potential to provide vital information to underpin successful production of pharmacologically active compounds, whether from roots or from seeds.

With the development of mapping populations for S. flavescens, it will be possible to investigate the genetic control of important traits. We are now in a strong position to make informed decisions about the most informative plants to select as parental combinations to maximise diversity. GBS could be applied to mapping populations and SNPs can readily be converted to KASP markers for high-throughput application in selective breeding. It would be highly desirable to have a complete genome sequence, as well as isoform sequencing. All of this is achievable with current technology, indicating that rapid progress is possible despite the constraints of working with a species that takes years to reach harvest potential.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

KC and JLi planned and executed the collection of seeds in China. KC established and maintained the field trial site. KC and DM developed the strategy for DArTSeq and KASP marker design and supervised its implementation. SNG prepared the DNA samples for DArTSeq, designed and assayed the KASP markers (designed from DArTSeq SNPs), and conducted the SNP based phylogenetic analysis. KC, RB, TB-M, and CS designed the strategy for analysis of QA genes, qPCR and QA metabolite analysis. KL performed the microscopy on the developing flower buds. CS obtained the SfCAO cDNA sequences (in silico and experimentally) and, performed and analysed the Sanger sequencing of SfLDC and SfCAO gene fragments. CS performed the purification and quantification of oxymatrine and matrine with significant protocol development assistance from JLa. CS, DM, and KC wrote the manuscript, with input from all authors. All authors have read and approved the final manuscript.

FUNDING

Funding was provided by Shanxi Zhendong Daodi Medicinal Materials Development Company Limited was not involved in the study design, collection, analysis, interpretation of data,
the writing of this article or the decision to submit it for publication. This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.

ACKNOWLEDGMENTS

We thank Rebecca Fox and Elysia Vassos for expert technical assistance with plants and molecular markers including the design of the gene specific KASP primers, and Sandy Khor for performing the qPCR.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.704201/full#supplementary-material

REFERENCES

Bunsupa, S., Hanada, K., Maruyama, A., Aoyagi, K., Komatsu, K., Ueno, H., et al. (2016). Molecular evolution and functional characterization of a bifunctional decarboxylase involved in Lycopodium alkaloid biosynthesis. Plant Physiol. 171, 2432–2444. doi: 10.1104/pp.16.00639

Bunsupa, S., Katayama, K., Ikeyura, E., Okawa, A., Toyooka, K., Saito, K., et al. (2012). Lysine decarboxylase catalyzes the first step of quinolizidine alkaloid biosynthesis and coevolved with alkaloid production in leguminosae. Plant Cell 24, 1202–1216. doi: 10.1105/tpc.112.095885

Burton, R. A., Collins, H. M., Kibble, N. A. J., Smith, J. A., Shirley, N. J., Jobling, S. A., et al. (2011). Over-expression of specific HvCellf cellulose synthase-like genes in transgenic barley increases the levels of cell wall (1,3;1,4)-β-D-glucans and alters their fine structure. Plant Biotechnol. J. 9, 117–135. doi: 10.1111/j.1467-7652.2010.00532.x

Carvajal-Larenas, F. E., Linnemann, A. R., Nout, M. J. R., Koziol, M., and van Boeckel, M. (2016). Lupinus mutabilis composition, uses, toxicology, and debittering. Crit. Rev. Food Sci. Nutrit. 56, 1454–1487. doi: 10.1080/10408398.2013.772089

Chen, L., Huang, X. B., Wang, H., Shao, J., Luo, Y., Zhao, K. R., et al. (2020). Integrated metabolomics and network pharmacology strategy for ascertaining the quality marker of flavonoids for Sophora flavescens. J. Pharm. Biomed. Analysis 186,113297. doi: 10.1016/j.jpba.2020.113297

Chen, X., Yi, C. Q., Yang, X. Q., and Wang, X. R. (2004). Liquid chromatography of active principles in Sophora flavescens root. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 812, 149–163. doi: 10.1016/S0170-2693(04)00679-8

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. G., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14, 1147–1159. doi: 10.1101/gr.1917404

Cui, J., Qu, Z. P., Harata-Lee, Y., Shen, H. Y., Aung, T. N., Wang, W., et al. (2020). The effect of compound kushen injection on cancer cells: integrated identification of candidate molecular mechanisms. PLoS One 15:e0236395. doi: 10.1371/journal.pone.0236395

Doyle, J. (1987). A rapid procedure for DNA purification from small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi: 10.1093/nar/gkh340

Han, R. C., Takahashi, H., Nakamura, M., Bunsupa, S., Yoshimoto, N., Yamamoto, H., et al. (2015). Transcriptome analysis of nine tissues to discover genes involved in the biosynthesis of active ingredients in Sophora flavescens. Biol. Pharm. Bull. 38, 876–883. doi: 10.1248/bpb.b14-00834

Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A., and Tilgner, H. U. (2019). Getting the entire message: progress in isoform sequencing. Front. Genet. 10:709. doi: 10.3389/fgene.2019.00709
Korbie, D. J., and Mattick, J. S. (2008). Touchdown PCR for increased specificity.

Kamphuis, L. G., Foley, R. C., Frick, K. M., Garg, G., and Singh, K. B. (2020). HKCMMS. (2012). A Wrapper Tool Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files.

Krueger, F. (2015). Marçais, G., Yorke, J. A., and Zimin, A. (2015). QuorUM: an error corrector for Illumina reads.

Kroc, M., Koczyk, G., Kamel, K. A., Czepiel, K., Fedorowicz-Strońska, O., Reumann, S., Buchwald, D., and Lingner, T. (2012). PredPlantPTS1: a web server for the prediction of plant peroxisomal proteins.

FigTree v1.4.2, A Graphical Viewer of Phylogenetic Trees. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed November 30, 2020).

Southern, E. M. (2000). Blotting at 25.

Tavladoraki, P., Cona, A., and Angelini, R. (2016). Copper-containing amine oxidases and FAD-dependent polyamine oxidases are key players in plant tissue differentiation and organ development. Front. Plant Sci. 7:824. doi: 10.3389/fpls.2016.00824

Tucker, S. (2003). Floral development in legumes. Plant Physiol. 131, 911–926. doi: 10.1104/pp.102.017549

Wang, P. H., Zhou, G. F., Jian, J. B., Yang, H., Renshaw, D., Aubeert, M. K., et al. (2021). Whole-genome assembly and resequencing reveal genomic imprint and key genes of rapid domestication in narrow-leaved lupin. Plant J. 105, 1192–1210. doi: 10.1111/tpj.15100

Wang, W., You, R. L., Qin, W. J., Hai, L. N., Fang, M. J., Huang, G. H., et al. (2015). Anti-tumor activities of active ingredients in compound kushen injection. Acta Pharmacol. Sin. 36, 676–679. doi: 10.1038/aps.2015.24

Wang, W. L., Xu, J. F., Fang, H. Y., Li, Z. J., and Li, M. H. (2020). Advances and challenges in medicinal plant breeding. Plant Sci. 298:110573. doi: 10.1016/j.plantsci.2020.110573

Wang, Y., Zhou, T. T., Li, D. H., Zhang, X. H., Yu, W. W., Cai, J. F., et al. (2019). The genetic diversity and population structure of Sophora alopecuroides (Faboideae) as determined by microsatellite markers developed from transcriptome. PLoS One 14:e0226100. doi: 10.1371/journal.pone.0226100

Wang, Y., Zhang, H., Chen, L. L., Sun, X. P., Yang, Q. J., Wan, L. L., and Guo, C. (2020). Matrine: a promising natural product with various pharmacological activities. Front. Pharmacol. 11:588. doi: 10.3389/fphar.2020.00588

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.