Transcriptomic study on two medicinal plants *Pueraria lobata* and *Sophora flavescens*

二種の薬用植物クズとクララにおける
トランスクリプトーム解析

---2015---

Rongchun Han

韓　栄春

Department of Molecular Biology and Biotechnology

Graduate School of Pharmaceutical Sciences

Chiba University

Table of Contents

List of Abbreviations

4CL	4-coumarate-CoA ligase
AC	accession number
BBB	Bud right before blossom
BLAST	Basic local alignment search tool
CA4H	*Trans*-cinnamate 4-monooxygenase
CHI	Chalcone isomerase
CHS	6'-deoxychalcone synthase
CuAO	Copper amine oxidase
DE	Differentially expressed
DRA	DDBJ sequence read archive
EC	Enzyme commission
EST	Expressed sequence tag
FDR	False discovery rate
GO	Gene ontology
HID	2-hydroxyisoflavanone dehydratase
IFS	2-hydroxyisoflavanone synthase
KEGG	Kyoto encyclopedia of genes and genomes
L/ODC	Lysine/ornithine decarboxylase
P450	cytochrome P450 monooxygenase
PAL	Phenylalanine ammonia-lyase
PBS	Pedicel while bud stage
P. lobata	*Pueraria lobata*
PWB	Pedicel while blossom
qRT-PCR	quantitative real-time reverse transcription PCR
RNA-Seq	Massively parallel cDNA sequencing
Root VC	Root vascular cylinder
RPKM	Reads per kilobase of transcript per million mapped reads
RT-PCR	Reverse transcription polymerase chain reaction
S. flavescens	*Sophora flavescens*
WEGO	Web gene ontology annotation plot

List of Tables

List of Figures

Abstract

Pueraria lobata (Willd.) Ohwi has a long and broad application in the treatment of disease. However, in the US and EU, it is treated as a notorious weed.

Sophora flavescens Aiton has long been used to treat various diseases. Although several research findings revealed the biosynthetic pathways of its characteristic chemical components as represented by matrine, insufficient analysis of transcriptome data hampered in-depth analysis of the underlying putative genes responsible for the biosynthesis of pharmaceutical chemical components.

The information to be gained from decoding the deep transcriptome profile would facilitate further research on *P. lobata* and *S. flavescens*. In this study, more than 93 million fastq format reads were generated by Illumina's next-generation sequencing approach using five types of *P. lobata* tissue, followed by CLC *de novo* assembly methods, ultimately yielding about 83,041 contigs in total. Then BLASTx similarity searches against the NCBI NR database and UniProtKB database were conducted. Once the duplicates among BLASTx hits were eliminated, ID mapping against the UniProt database was conducted online to retrieve Gene Ontology information. In search of the putative genes relevant to essential biosynthesis pathways, all 1,348 unique enzyme commission numbers were used to map pathways against the Kyoto Encyclopedia of Genes and Genomes. Enzymes related to the isoflavonoid and flavonoid biosynthesis pathways were focused for detailed investigation and subsequently,

qRT-PCR was conducted for biological validation. Metabolites of interest, puerarin and daidzin were studied by HPLC.

For *S. flavescens*, more than 200 million fastq format reads were generated by next-generation sequencing approach using its nine types of tissue, CLC de novo assembly produced 83,325 contigs over 300 bp. RPKM values were calculated to analyze gene expression levels, and overrepresented gene ontology terms were evaluated using Fisher's exact test. To study its characteristic metabolic pathways, all 1,350 unique enzyme commission numbers of *S. flavescens* were used to map pathways against KEGG. The preferential expression of the gene for putative lysine/ornithine decarboxylase committed in the initial step of matrine biosynthesis in leaves and stems was confirmed in semi-quantitative PCR analysis. By analyzing expression patterns, we proposed some candidate genes involved in the biosynthesis of isoflavonoids and quinolizidine alkaloids.

Adopting RNA-Seq analysis, we obtained substantially credible contigs for downstream work and the findings in this report may serve as a stepping-stone for further research into the promising leguminous medicinal plants.

General Introduction

Ⅰ Studies Fueled by Phytochemical Genomics

Phytochemical genomics is a recently emerging filed, which investigates the genomic basis of the synthesis and function of phytochemicals (plant metabolites) (Saito, 2013). Central dogma of molecular biology described the sequential information flow in biological systems: through replication, DNA could be copied to produce DNA and then mRNA can be achieved by transcription, followed by translation to synthesize proteins using mRNA as a template. A portion of proteins will work as enzymes to catalyze the reactions *in vivo* in various organisms to generate metabolites. In order to study the biosynthetic mechanism and regulation, function and evolution of plant metabolites, systematic integration of genomics and related '-omics' such as transcriptomics, proteomics and metabolomics will be very useful (Saito, 2013).

Hypotheses regarding the related scientific domains can be generated and tested by this integrated systematic analysis. One of the many powerful approaches is taking advantage of integrated functional genomics. By studying the correlation of co-responding elements in transcriptome, the related genes responsible for the production of certain metabolites can be grouped together for gene expression and metabolite accumulation study.

Statistical analysis suggested there are up to 1 million plant metabolites and phytochemical genomics may hold the key to how these specialized plant products are produced and regulated (Muranaka and Saito, 2013).

Ⅱ Advent and Rapid Development of Next-generation Sequencing

Next-generation sequencing (NGS) is the newer approach compared to the automated Sanger's method which is regarded as a 'first-generation' sequencing approach. During the first-generation era, there were several well-established sequencing techniques, such as the Sanger method, DNA sequencing in real time by the detection of the released pyrophosphate (also known as the Pyrosequencing method), the Maxam and Gilbert method, and single molecule sequencing, which showed different features (Franca, et al., 2002). The automated Sanger method was popular and widely used for almost two decades and resulted in a series of great accomplishments, including the achievement of the human genome project.

As a part of NGS, RNA-Seq utilizes newly developed deep-sequencing techniques. Total or fractionated RNA is converted to a library of cDNA. After fragmentation, adaptors are attached to each cDNA fragment at one or both ends. Using high-throughput sequencing instrumentation, short (usually 30-400bp) raw reads from one end or both ends will be obtained (**Figure** Ⅰ) (Wang et al., 2009). The resultant sequence reads can then be aligned to the reference genome or transcriptome, or using *de novo* assembly approaches to generate contigs or unigenes.

In the early 1990s, the cost for DNA sequencing would be expected at about $0.12 to $0.15 per base (Adams et al., 1991). However, 20 years later, the cost was reduced rapidly to $0.07 per million bases (Liu et al., 2012). In addition to the drastic cost reduction, time span, accuracy and throughput for the whole process have also been improved significantly.

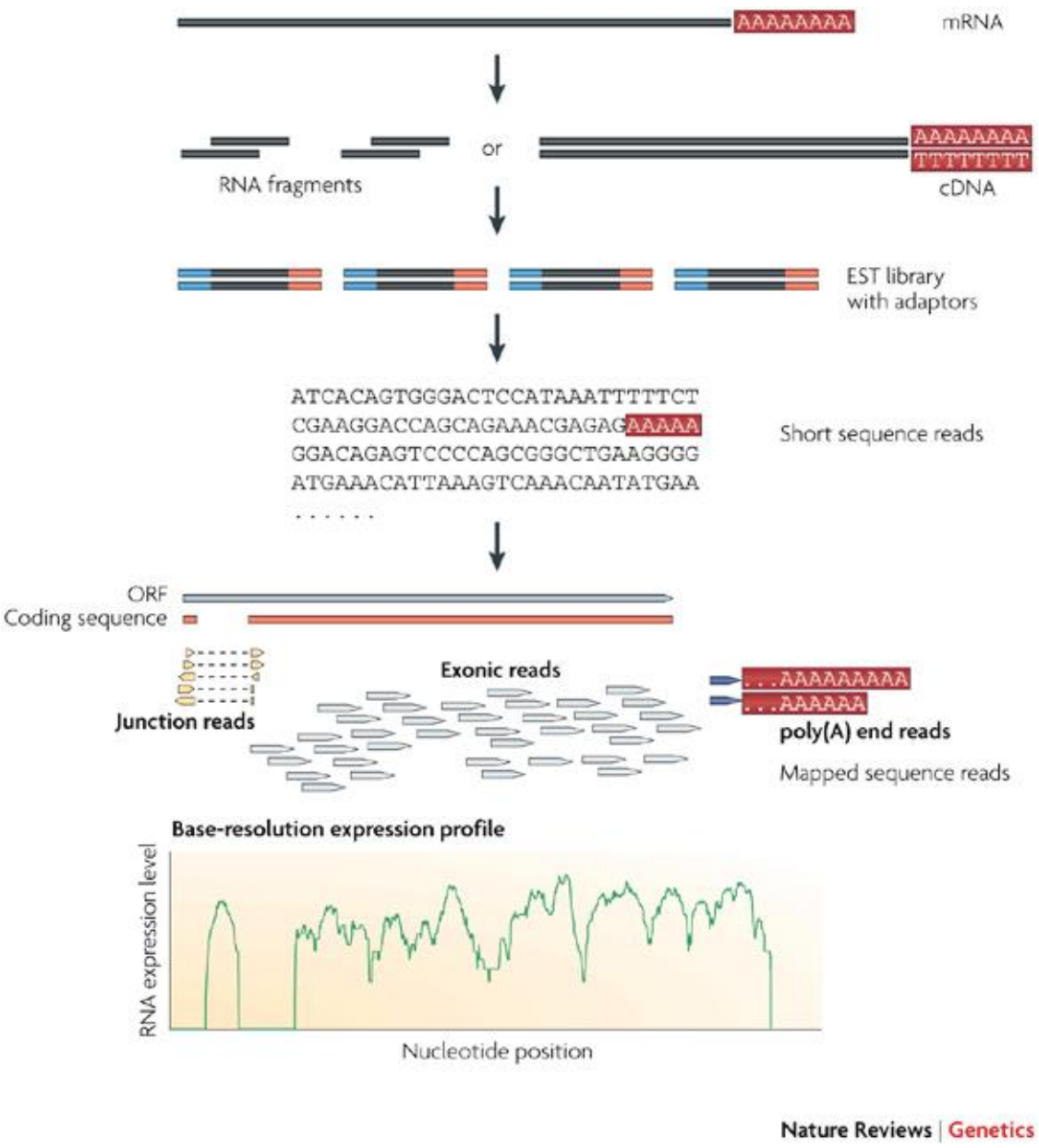

Figure I | A typical RNA-Seq experiment

Ⅲ Usage of *de novo* deep transcriptome assembly

Typically, for studying model plants for which genome information is available, to map the raw reads back to the plants' genome would not be a big challenge. While the unavailability of the genome hampers the research for the non-model plants.

With the rapid development of NGS, several practical *de novo* assembly platforms emerged to deal with tremendous amount of the information obtained by deep sequencing technologies, including Trinity (Grabherr et al., 2011), CLC (CLC bio, Aarhus, Denmark), Assembly by Short Sequences (AbySS) (Simpson et al., 2009), Velvet (Zerbino and Birney, 2008), Short Oligonucleotide Analysis Package (SOAPdenovo) (Li et al., 2009).

For *de novo* assembly, a table of all sub-sequences of length k (k-mers) is generated from every single read and length k may vary in different experiments to achieve the best performance.

By using de Bruijn graph, all the potential neighboring sub-sequences will be considered to extend the assembled contigs if shifting a k-mer by one character creates an exact k-1 overlap between the two k-mers. **Figure Ⅱ** (Martin and Wang, 2011) shows an example of 5-mers. Due to sequencing error or SNP, the resultant de Brujin graph may consist of 'bubbles'. Then the chains are merged and the simplified graph with the existing bubbles is achieved. Finally, consider all the possible alternative paths in each assembled de Brujin graph and generate the isoforms (Martin and Wang, 2011).

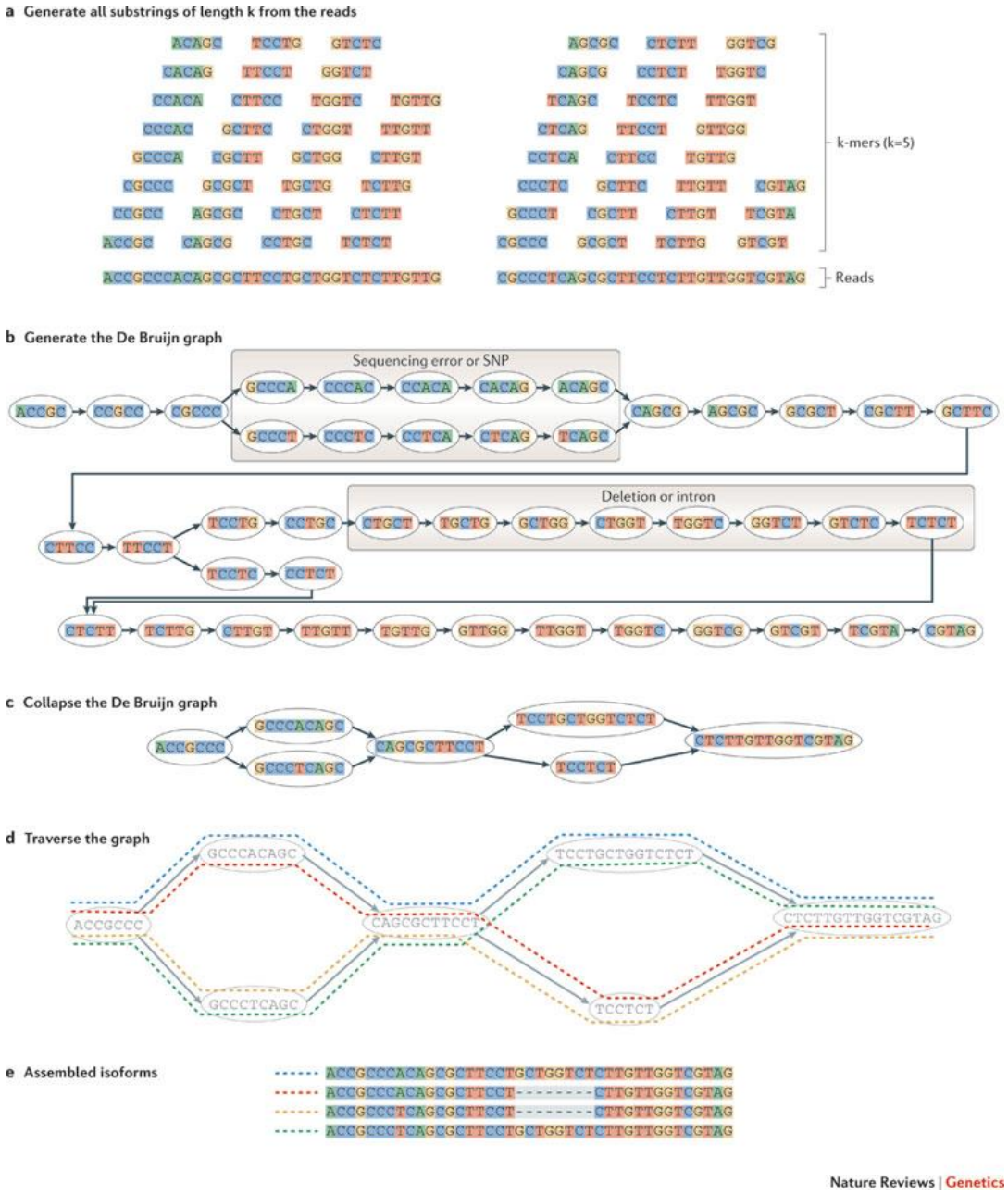

Figure Ⅱ| Overview of *de novo* deep transcriptome assembly

Ⅳ Flavonoids and Isoflavonoids Biosynthesis in Plants

As phenolic secondary metabolites, isoflavones are mainly found in leguminous plants. Studies have shown that isoflavones are involved in plenty of plat-microbe interactions.

Both flavonoids and isoflavonoids are accumulated as major plant secondary metabolites which demonstrate various biological features as well as show important ecological impacts. Such constituents not only take part in critical physiological processes but also exert beneficial effects on human health, ranging from decreasing cholesterol levels and prevention of certain solid tumors to improving women's health (Ralston et al., 2005).

Followed by phenylalanine biosynthesis, the essential amino acid is converted to cinnamic acid with the help of phenylalanine ammonia-lyase (PAL). Subsequently, trans-cinnamate 4-monooxygenase (CA4H) play the role to generate 4-coumaric acid. 4-coumarate-CoA ligase (4CL) catalyzes the reaction to produce 4-coumaroyl CoA and by far, the acquired compounds can be used as precursors in various biological processes.

The dedicated part responsible for the biosynthesis of flavones starts from chalcone synthase (CHS) catalyzing the reaction to obtain chalcone scaffolds. For the following steps to produce flavonoids and isoflavonoids, although the typical procedure is conserved in plants, depending on the species, many enzymes including isomerases, reductases, hydroxylases and *C*- or *O*-glycosyltransferases kick in to modify the structures and modulate the physiological activity by changing the solubility and reactivity (Maria et al., 2012).

Ⅴ Transcriptomic Analysis on *P. lobata* and *S. flavescens*

In spite of the powerful methods recently developed including phytochemical genomics and the protruding urgency of applying RNA-Seq analysis on medicinal plants, the unavailability of deep transcriptomic data for the two leguminous plant *P. lobata* and *S. flavescens* hinders their in-depth study.

P. lobata demonstrates potent efficacy in treating diseases such as alcoholism (Carai et al., 2000) and diabetic retinopathy (Teng et al., 2009; Cherdshewasart et al., 2007). However, since it was introduced to the United States of American in the 19th century and then to the European countries, due to its aggressive growth rate, it has long been regarded as a major biosystem threat. The general non-targeted deep transcriptomic analysis would shed light into its underlying mechanism and provide clues to solve the problem in an ecology perspective.

S. flavescens has been recorded and used for more than 1,800 years (Sun et al., 2012). As a widely distributed and effective herbal medicine, it severed as a cure for asthma, sores, gastrointestinal hemorrhage, diarrhea, allergy, inflammation in eastern Asian countries (Hong et al., 2009; Funaya and Haginaka, 2012). Main chemical components of *S. flavescens* include flavonoids (1.5%), alkaloids (3.3%), alkylxanthones, quinones, triterpene glycosides, fatty acids as well as essential oils and recently, several clinical studies reported that alkaloids of *S. flavescens* were efficacious in treating various types of solid tumors (including breast, lung, liver and gastrointestinal tract cancers), which drew close attention to this traditional

herbal plant (Sun et al., 2012; Li et al., 2012). RNA-Seq analysis on the multiple tissues of this plant will deepen our understanding concerning its overall biological nature and specialized biosynthetic pathways as well as metabolite accumulation.

Ⅵ Thesis Work

In this study, five tissues of *P. lobata*, leaf, mature root, root vascular cylinder, young root and stem were collected for RNA-Seq analysis. We assembled all the contigs and uploaded the raw reads to the public repository Kyoto Encyclopedia of Genes and Genomes (KEGG) for the overall better understanding of this medicinal plant. Because the multiple tissues bear distinct accumulation profile of various characteristic compounds, we employed the method to calculate RPKM values and furthermore assessed the differentially expressed transcripts.

The characteristic compound found in *P. lobata* is puerarin. Researchers have tried for a long period of time to puzzle out its biosynthesis. Although the biosynthetic pathway for flavonoids and some of the isoflavonoids has been well established, the enzyme responsible for the formation of this special *C*-glycoside remains a mystery. By investigating the co-expression information, we proposed a list of candidate glucosyltransferases for future research. As for the isoflavonoid biosynthetic pathway in *P. lobata*, candidate enzymes were picked out and subjected to qRT-PCR for biological validation.

For *S. flavescens*, nine tissues (callus, leaf, flower, stem, young bud,

mature bud, bud right before blossom, pedicel while bud stage and pedicel while blossom) were sampled for next generation sequencing. In order to study the biosynthesis of its major quinolizidine alkaloids including matrine and oxymatrine, based on the knowledge currently available, co-expression analysis was performed to find the closely related transcripts to the identified gene lysine/ornithine decarboxylase (*L/ODC*) and semi-quantitative PCR was conducted to verify the expression profile of *L/ODC* in different tissues of *S. flavescens*.

This manuscript is divided into two chapters detailing the findings acquired throughout the whole thesis work. Chapter one covers the study on different aspects of *P. lobata* deep transcriptome data. Chapter two summarizes the experiment regarding RNA-Seq analysis on *S. flavescens*.

Chapter 1: Transcriptomic Landscape of *Pueraria lobata* Demonstrates Potential for Phytochemical Study

1.1 Introduction

Pueraria lobata (Willd.) Ohwi (Kudzu) has been described and used as a traditional medicinal plant for more than 20 centuries in oriental medicine (Keung and Vallee, 1998). The *P. lobata* root, a part of the plant that is prescribed most frequently, accumulates abundant polyphenolic compounds, including isoflavones, isoflavonoid glycosides, coumarins, puerarols and the associated derivatives (Wong et al., 2011). Intensive investigation has revealed a chemical profile with antioxidant and antimutagenic activity (Miyazawa et al., 2001; Cherdshewasart and Sutjit, 2008) and efficacy in the treatment of alcoholism (Carai et al., 2000) and diabetic retinopathy (Teng et al., 2009; Cherdshewasart et al., 2007). The plant was introduced to the United States in 1876 as an ornamental plant and then to Europe. Due to its rapid growth and vigorous adaptation to the surroundings, *P. lobata* is now regarded as a major ecosystem threat (Follak, 2011) and a noxious weed, according to the USDA plant database. In order to evaluate the plant's potential as a cure for disease or regulate its invasive influence on other native plants, additional research using next-generation sequencing technologies into this leguminous plant is necessary.

The genomes for many model organisms have been sequenced but for these non-model plants, the lack of reference genome information hinders

studies on the underlying genes involved in essential biological processes related to drug development. In this regard, transcriptomic sequencing plays an essential role in understanding the genetic diversity across organisms. Such approaches elucidate the genetic code that underlies protein diversity (Muranaka and Saito, 2013; Saito, 2013). The use of new technologies such as the Short Oligonucleotide Analysis Package (SOAPdenovo) (Li et al., 2009), Assembly by Short Sequences (AbySS) (Simpson et al., 2009), and Trinity (Grabherr et al., 2011) accelerates the pace of transcriptomic profiling when processing tremendous amounts of data generated from large-scale sequencing projects. Massively parallel cDNA sequencing (RNA-Seq) measures the levels of transcripts and their isoforms far more precisely than other methods (Fullwood et al., 2009; Wang et al., 2009).

C-glycosides are widespread in plants, insects and microbes, where they serve a diverse range of functions including acting as antibiotics, antioxidants, attractants and feeding deterrents (Brazier-Hicks et al., 2009). Early study on *Fagopyrum esculentum* seedlings showed 2-hydroxylation of flavanones was a critical prerequisite for the corresponding *C*-glucosyltransferase to catalyze (Kerscher and Franz, 1987). Recently, reports regarding *C*-glycosylation of flavonoids in crops such as wheat and corn suggested considerable similarity of the proteins some of which exhibited bifunctional *C*- and *O*-glucosyltransferase activity (Ferreyra et al., 2013). For the characteristic compound daidzein-8-*C*-glycoside (puerarin) found in *P. lobata*, although the biosynthetic pathway for daidzein in

legumes is well established (Steele et al., 1999; Jung et al., 2000), the key enzyme responsible for catalyzing this isoflavone aglycon remains to be identified.

In this study, 93,248,914 paired-end reads of *P. lobata* were generated from five different tissues by Illumina's sequencing platform. Illumina reads were deposited at the DDBJ Sequence Read Archive (DRA) with accession number DRA001736 and the resultant contigs along with the top hits of BLASTx at GitHub (https://github.com/rongchunhan/Pueraria_lobata). CLC Genomics Workbench (CLC bio, Aarhus, Denmark) was subsequently applied to conduct *de novo* assembly. Based on the findings provided by Gene Ontology and KEGG pathway mapping, the candidate genes that may be involved in the biosynthesis of key chemical components were identified. For biological validation, quantitative real-time reverse transcription polymerase chain reaction (qRT-PCR) was applied to check the genuine expression profile for the genes involved in the biosynthetic pathway leading to isoflavonoids. Meanwhile, the concentrations of puerarin and daidzin, the characteristic compounds in Kudzu, were measured by High Performance Liquid Chromatography (HPLC) within the five tissues from which we obtained the deep transcriptomic data.

1.2 Materials and Methods

1.2.1 Plant Materials, Chemicals and Total RNA Extraction

Fresh tissues and organs were collected from healthy *P. lobata* plants growing in Chiba, Japan in May 2012. Puerarin and daidzin standard substances were purchased from LC laboratories (USA). The materials were dipped into RNA stabilization solution (RNA*later*, Life technologies, USA) immediately after removal from the field. The RNA*later* solution was gently removed with a Kimwipe.

Then the remaining sample was frozen by liquid nitrogen and powdered using Multi Beads Shocker (Yasui Kikai, Japan). TRIzol Reagent (Invitrogen, USA) was used to extract total RNA from powdered *P. lobata*. The RNA obtained was then treated using the RNeasy Mini Kit (Qiagen, USA).

1.2.2 cDNA Library Preparation and Sequencing

The TruSeq RNA Sample Prep Kit v2 (Illumina, CA, USA) was used for cDNA library preparation and sequencing. Once the mRNA in total RNA had been polyA-selected and fragmented, double-stranded cDNA was prepared for cDNA library construction. After the creation of blunt-end fragments and indexed adaptor ligation, the samples were hybridized to flow cells.

Cluster amplification was completed using the cBot Cluster Generation System (Illumina, CA, USA) and then sequenced by Illumina's next-generation sequencing instrument, the HiSeq 1000 (Kozarewa et al.,

2009).

1.2.3 CLC Approach to *de novo* Assembly

Prior to assembly, the original fastq *P. lobata* format data were subjected to CLC trimming to eliminate reads of poor quality. The CLC method (version 4.9) was used to process these clean reads. The publically available *P. lobata* expressed sequence tags (ESTs) data (6,365) were downloaded from the National Center for Biotechnology Information (NCBI) database of expressed sequence tags (dbEST).

All resultant contigs over 200 bp were taken into consideration for the downstream analysis. Because the assembly process may result in duplicate contigs, CD-HIT-EST was applied with representative sequences at 90% identity to obtain unique unigenes (Fu et al., 2012).

1.2.4 Transcript Abundance and Expression-based Analysis

RPKM stands for reads per kilobase of the transcript per million mapped reads. The formula is as follows: $\text{RPKM} = \frac{10^9 \times C}{N \times L}$, where C is the number of reads mapped to the gene's exons, N the total number of mapped reads in the experiment and L the total length of the exons in base pairs.

In order to estimate contig expression level, we applied the CLC approach to map all fastq format reads back to the contigs and calculated the RPKM values. Because the five samples lacked technical replicates, the non-parametric approach for the identification of differentially expressed genes, NOISeq-sim (Tarazona et al., 2011), was adopted to analyze ten

independent pair-wise sample comparisons.

1.2.5 Annotation Pipeline and Data Mining

All *de novo* contigs were used as query sequences for the BLASTx sequence similarity search against the non-redundant (NR) protein database at NCBI and the Universal Protein resource (UniProt) at UniProt consortium (Magrane and Consortium, 2011).

The e-value threshold was set to 1e−10; the upper limit on the number of subject sequences from databases to show alignment was limited to 20. As to the large BLASTx output, only percent identities over 40% and e-values less than 1e−30 were taken into consideration.

After eliminating redundancies, all unique gene identifiers in fasta format were then uploaded to the UniProt ID mapping website for online data processing (http://www.uniprot.org).

By consolidating the returned target list and the UniProtKB accession numbers (ACs) obtained from the above-mentioned BLASTx output against the UniProt database, we applied the redundancy-free ACs to annotation using the same online facilities. Out of the huge number of annotation results, we examined the reviewed findings from UniProtKB/Swiss-Prot as well as TrEMBL for data mining. The sequences with Gene Ontology (GO) terms at the protein level were classified. Ultimately, 1,348 enzyme commission (EC) numbers were applied to map pathways against KEGG, and the enzymes related to daidzein biosynthesis were depicted.

1.2.6 Gene Expression Validation Adopting qRT-PCR

qRT-PCR was conducted using 96-well plates and StepOnePlus real-time PCR system (Applied biosystems). Three technical replicates were used for each reaction, and a negative control consisting of template without primers was included for each template.

Reaction volume was 15 μL and each reaction comprised 7.5 μL of SYBR Select Master mix (2×), 0.15 μL of 10 μM primers (1:1 mix of forward and reverse primers), 1.0 μL of cDNA synthesized using SuperScript VILO cDNA Synthesis Kit (Life technologies), and 6.35 μL of nuclease-free distilled water.

Reaction conditions included 10 min incubation at 95℃, then 40 cycles of 95℃ for 15 sec and 60℃ for 1 min, followed by a melt-curve analysis to confirm single PCR product amplification. β-actin was used as the internal control gene(Hong et al., 2010). No amplification was observed in any negative control.

Equivalent slopes for target and internal control gene were observed in amplification plots, so the comparative threshold-cycle (C_T) method was used to calculate relative expression levels as $2^{-\Delta Ct}$ where ΔCt = (C_T target gene – C_T internal control gene), assuming similar PCR efficiencies of target and internal control gene (Schmittgen and Livak, 2008; Gaines et al., 2014).

1.2.7 HPLC Analysis of Puerarin and Daidzin in 5 Tissues of *P. lobata*

1 g of every fresh tissue studied was ground to powder in liquid nitrogen and extracted overnight with 5 ml of acetone at 4℃. Then the extract was centrifuged at 3,500 rpm for 30 min (He et al., 2008; He et al., 2011) and the supernatant was dried in ventilator.

The residues were resuspended in methanol, and 10 μL of the solution was analyzed by reverse-phase HPLC (HITACHI D7000 system) on a 5-μm C15 column (150 × 4.6 mm, Mightysil). The flow rate was 0.5 mL/min with the mobile phase methanol/water (25:75).

1.3 Results

1.3.1 Plant RNA Extraction and cDNA Library Preparation

Studies on legumes showed the isoflavones daidzein and genistein were major metabolites in all embryonic organs within the dry seeds. Seedling roots and callus cultures are known to produce daidzein, with the highest daidzein concentration to be found in mature fruits (Graham, 1991; Bouque et al., 1998). We aimed to collect information about the nature of the genes responsible for the biosynthesis of daidzein and daidzin in *P. lobata*. We extracted total RNA from the leaf, mature root, root vascular cylinder (Root VC), young root and stem of the plant. Five distinct cDNA libraries were established from these five tissue samples. We will refer to the five libraries in the following manner: Library 1 (leaf), Library 2 (mature root), Library 3 (root VC), Library 4 (young root), Library 5 (stem). Wherever applicable, a uniform color scheme will be used to represent the libraries: red (Library 1), purple (Library 2), green (Library 3), blue (Library 4) and yellow (Library 5).

1.3.2 Illumina Sequencing and *de novo* Assembly

All five libraries were processed using the Illumina HiSeq 1000 platform. Empty reads, reads of low quality and those containing unknown bases were trimmed using CLC software. In order to consolidate the available bio-information to obtain more reliable and thorough findings, we combined the resultant clean reads with *P. lobata* EST sequences obtained from the NCBI database to conduct *de novo* assembly; thus, 83,041 contigs

were generated. By applying CD-HIT-EST with a threshold of 0.9, duplicates were identified and removed, leaving 81,508 non-redundant contigs for downstream analysis. An overview of the experimental pipeline is shown in **Figure 1.1**. **Table 1.1** and **Table 1.2** summarize trimming, sequencing and assembly results.

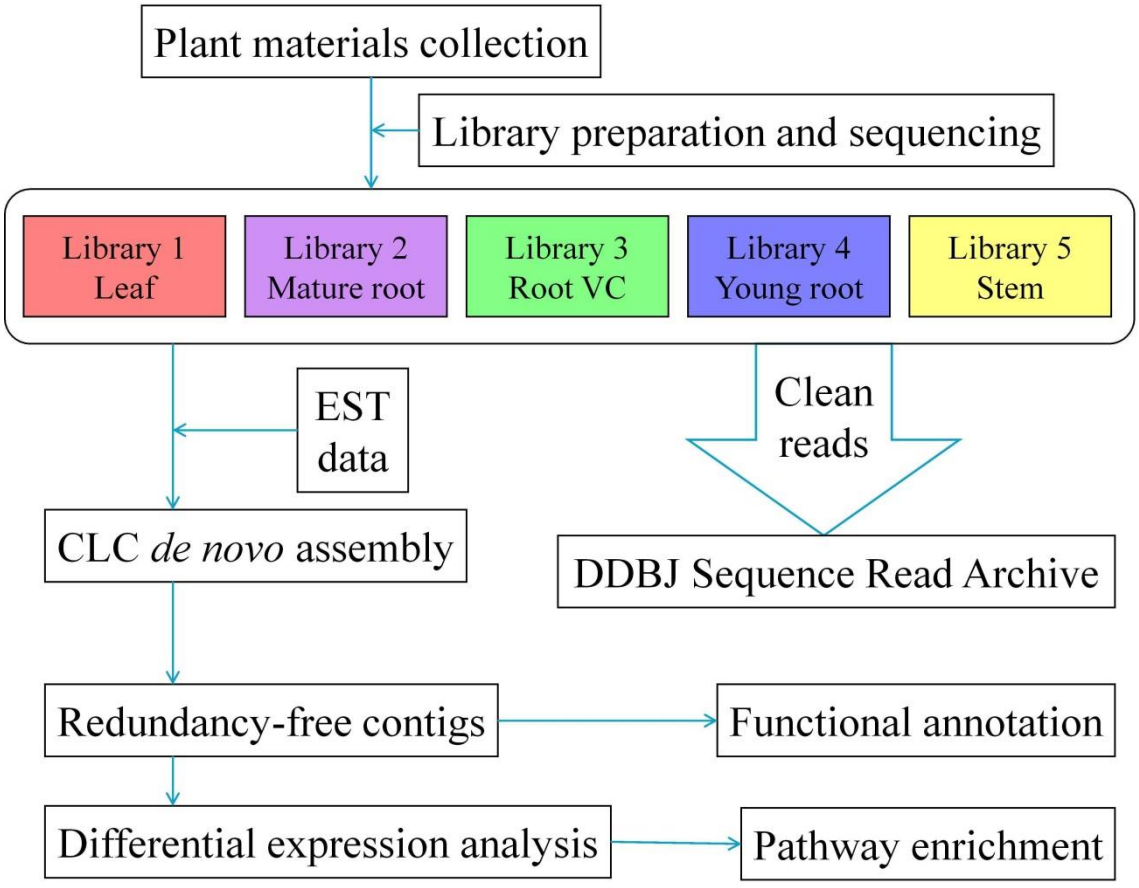

Figure 1.1 | Summary of the experimental design and analysis pipeline.

Table 1.1 Quality control for transcriptomic analysis

Library	No. of reads (paired)	Average length (bp)	No. of reads after trim (bp)	Average length after trim (bp)
1 (Leaf)	18,247,136	101.0	18,189,752	98.8
2 (Mature root)	14,090,648	101.0	14,039,985	98.2
3 (Root VC)	20,847,896	101.0	20,776,909	98.7
4 (Young root)	23,941,990	101.0	23,869,723	98.8
5 (Stem)	16,429,912	101.0	16,372,545	98.5

Table 1.2 Overview of *P. lobata* transcriptomic assembly

Items	Numbers
Total bases	9,197,584,658
Average length of reads (bp)	98.6
No. of reads (6,365 ESTs included)	93,255,279
Average length of contigs (bp)	730
N75; N50; N25 (bp)	488; 1,145; 2,125
No. of contigs over 200 bp	83,041
Non-redundant contigs	81,508

1.3.3 Guanine-Cytosine (GC) Content Analysis

The reported GC content for unigene sequences in soybean and *Arabidopsis* is 0.43 and 0.44, respectively (Tian et al., 2004; Kawaguchi and Bailey-Serres, 2005). The average GC content of *P. lobata* transcripts was found to be 39.9% (**Figure 1.2**). In eukaryotes, mean GC content varies from ~20 to 60% (Serres-Giardi et al., 2012). Our values are in the

middle of this range, slightly lower than those reported for *Glycine max* (43%) but very close to those reported for *Medicago truncatula* (40%) (Tian et al., 2004).

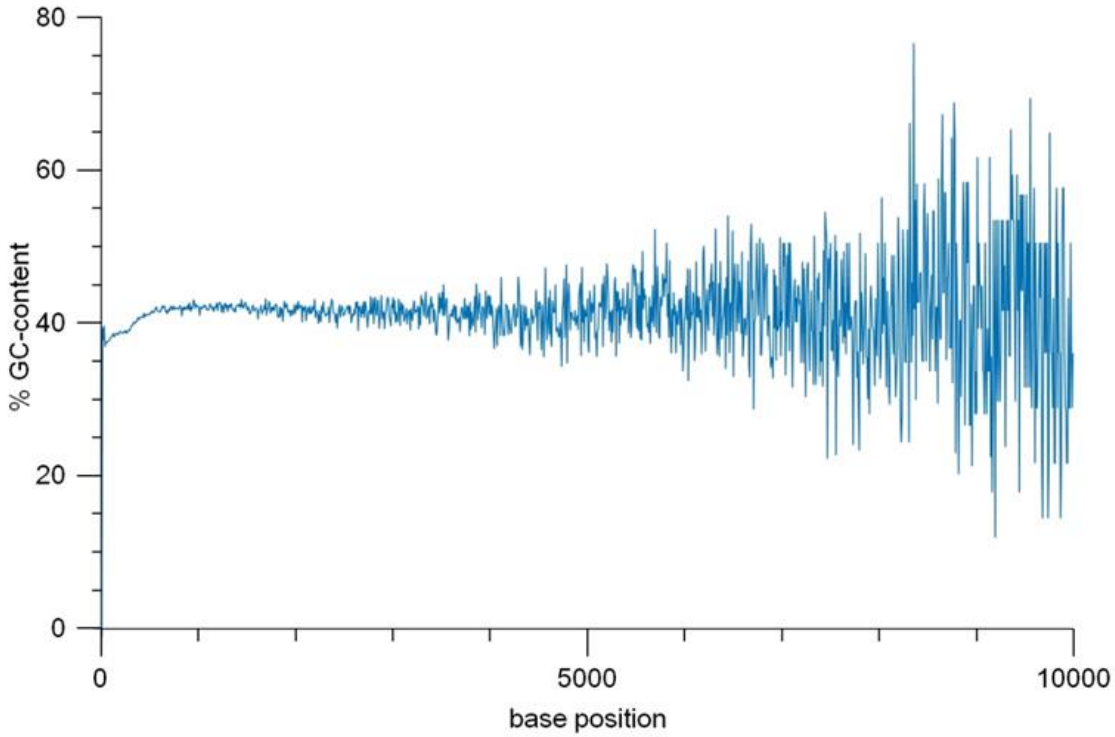

Figure 1.2 | GC content at different base positions for all *P. lobata* contigs. The length of the contigs varies from 200 to 15,631 bp. However, the GC content presented here does not calculate contigs information for more than 10,000 bp.

1.3.4 Transcriptome Information and Differential Accumulation of Transcripts

RPKM calculation was performed as the first step of transcript expression analysis after RNA-seq reads from every library were aligned to all contigs. A Venn diagram was drawn by utilizing the R project in conjunction with the Venn Diagram package (Chen and Boutros, 2011) to illustrate the distribution profile of all active contigs (78,201) with RPKM values >0 in at least one of the libraries (**Figure 1.3**).

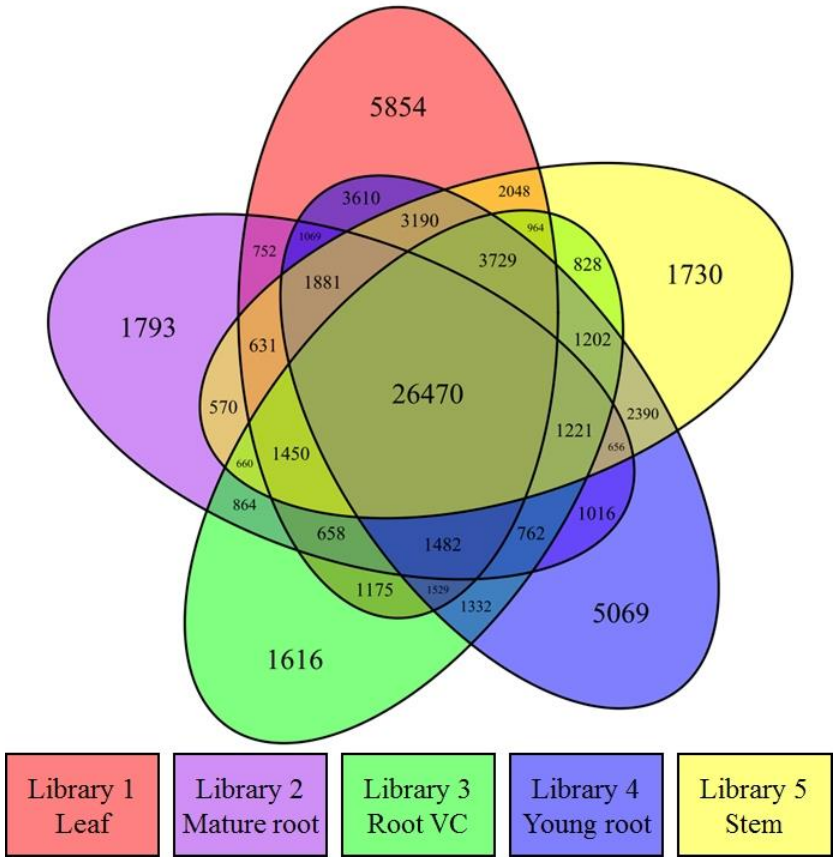

Figure 1.3 | Transcriptomic expression analysis. A Venn diagram shows the distribution of transcriptionally active contigs whose RPKM values are greater than 0 in at least one of the libraries.

33.8% of the active transcripts were expressed across all 5 libraries, suggesting the homogeneity and high quality of the acquired raw data. Compared to other three tissues, young root and leaf had more exclusively expressed contigs, which demonstrated in spring, such tissues played important and unique physiological roles.

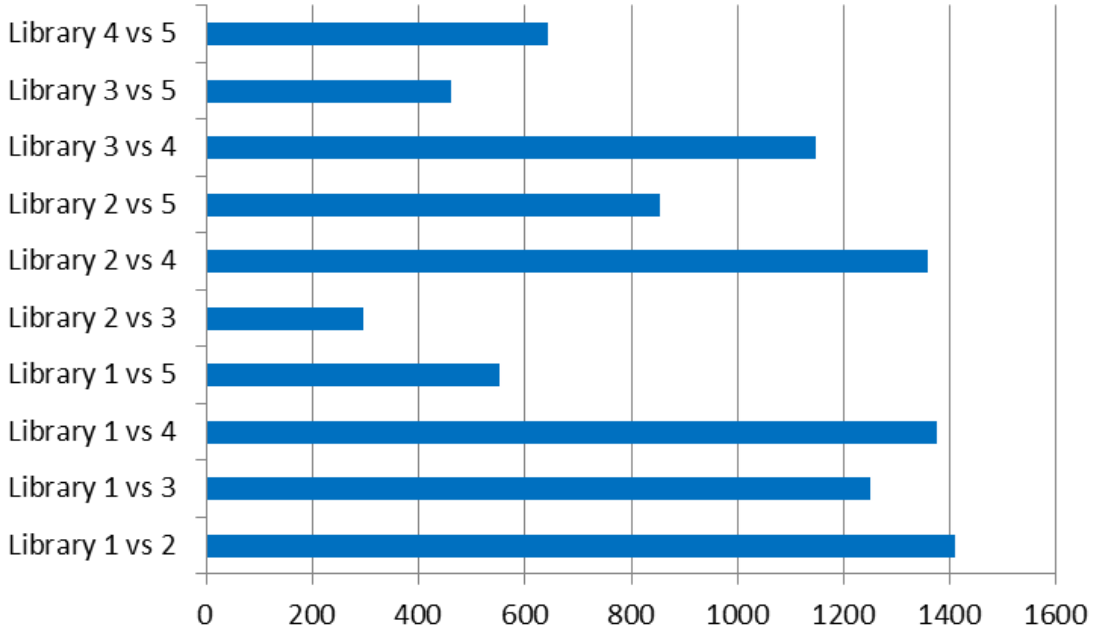

Figure 1.4 | Number of differentially expressed transcripts determined by NOISeq-sim for pairwise comparisons among the 5 libraries.

RPKM calculation considers gene length variation and the number of total mapped reads, which allows this normalized output to be used directly for the comparison of gene expression. To perform differential expression

(DE) analysis, ten pair-wise comparisons for the five libraries were conducted by applying NOISeq-sim, which is a non-parametric approach for the identification of differentially expressed genes from count data or previously normalized count data. By running NOISeq-sim on an R language platform with the given threshold (q = 0.9) for selecting differentially expressed features, the resultant number of DE transcripts varied across comparisons. The highest value obtained was 1,408 differences between leaf and mature root transcripts; the lowest value obtained was 297 differences between mature root and root vascular cylinder (**Figure 1.4**).

1.3.5 Protein Function Annotations and Gene Ontology (GO) Classification

Functional annotations according to sequence similarity are often the initial step in studying the role and biological functions of gene products (Ramilowski et al., 2013).

The basic local alignment search tool (BLAST) was utilized to scan nucleotide query sequences against protein databases (NR, UniProt) to identify similar subject sequences. When the threshold e-value for BLASTx searches was set to 1e−10 and the top 20 subject sequences for each query sequence were taken into consideration, we obtained 829,087 subject sequences for all 81,508 query sequences. To obtain reliable results while reducing redundancy, we set stricter requirements for retrieving the candidate genes. With this approach, significant matches were assigned to 30,156 contigs.

Gene Ontology, consisting of three main domains (cellular components, molecular functions and biological processes), is a useful instrument with which to study the nature of annotated genes (Ashburner et al., 2000). Based on NCBI NR BLAST results, with the aid of Web Gene Ontology Annotation Plot (WEGO) software (Ye et al., 2006), 26,245 contigs yielded corresponding GO terms that could be further classified into 48 sub-categories: 12 related to cellular components, 13 to molecular function and 23 to biological processes.

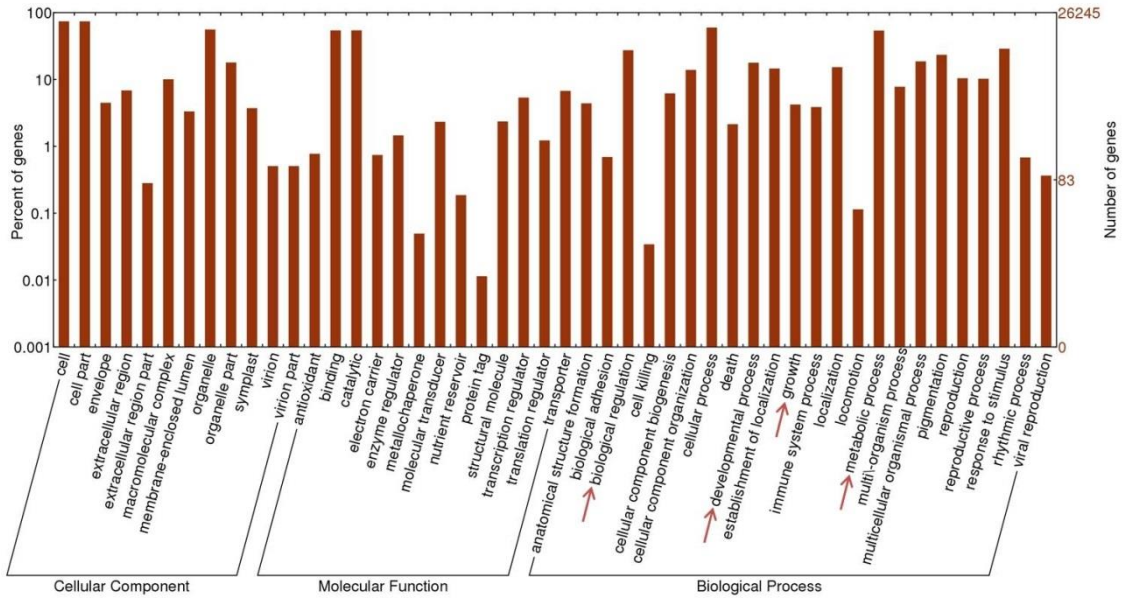

Figure 1.5 | Gene Ontology annotation for *P. lobata* contigs.

48 subcategories are affiliated to three main domains: molecular function, cellular components and biological processes.

Figure 1.5 presents the large number of transcripts related to metabolic processes (14,164) and biological regulation (7,194). The contigs assigned to "growth" and "developmental process" under GO biological process may hold the key to the mechanism underlying its rapid and aggressive growth rate.

The disadvantage of evaluating gene classification by directly counting the number of GO terms which possess the same or very similar functions is that the expression level of these query sequences varies, which grants distinct weight to the same GO term as it corresponds to different query sequences.

With this concern, overrepresented GO terms were identified by Fisher's exact test. The one-tailed Fisher's exact *p*-values corresponding to overrepresented categories were calculated according to the counts in 2×2 contingency tables. Counts n_{11}, n_{12}, n_{21} and n_{22} in each table stand for: n_{11}, number of observations of a specific category in the first gene set; n_{12}, number of other categories in the first gene set; n_{21}, number of observations of a category in the second gene set; and n_{22}, number of observations of other categories in the second gene set (Takahashi et al., 2011). *P*-values were corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) with the threshold set at 0.05. For each *P. lobata* library, contigs with RPKM value over 15.0 (the top ~10% of all transcripts) were regarded as highly expressed genes and extracted respectively. Then the merged 14,364 contigs were used to perform Fisher's exact test.

As with many woody vines supported by trees or man-made structures, Kudzu may allocate the majority of its biomass to vine elongation and leaf growth. It is regarded as an invasive alien plant in Europe and northern America because its rapid growth rate (up to 30 cm d^{-1}) (Lindgren et al., 2013) allows Kudzu to suffocate neighboring plants that are deprived of sunlight. The overrepresented GO terms (**Supplementary 1.1**, selected GO terms with p < 1E−30) suggest that highly activated biological processes such as cell division (GO: 0051301), cell growth (GO: 0016049), root hair elongation (GO: 0048767), response to cold (GO: 0009409) and response to salt stress (GO: 0009651) could play an essential role in its aggressive development.

1.3.6 KEGG Pathway Retrieval

The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012) provides a robust instrument for biological pathway assignment as well as the functional annotation of gene products. Based on ID mapping results, we obtained 1,348 unique enzymes corresponding to 16,380 contigs and subsequently retrieved pathways using KEGG. These EC numbers were assigned to 152 biological pathways with the largest number of enzymes (697) involved in metabolic pathways. Given the remarkable reputation of legumes with regard to the ability to accumulate functional flavonoids, 19 flavonoid biosynthetic and 14 isoflavonoid biosynthetic enzymes are presented in **Figure 1.6** and **Figure 1.7**.

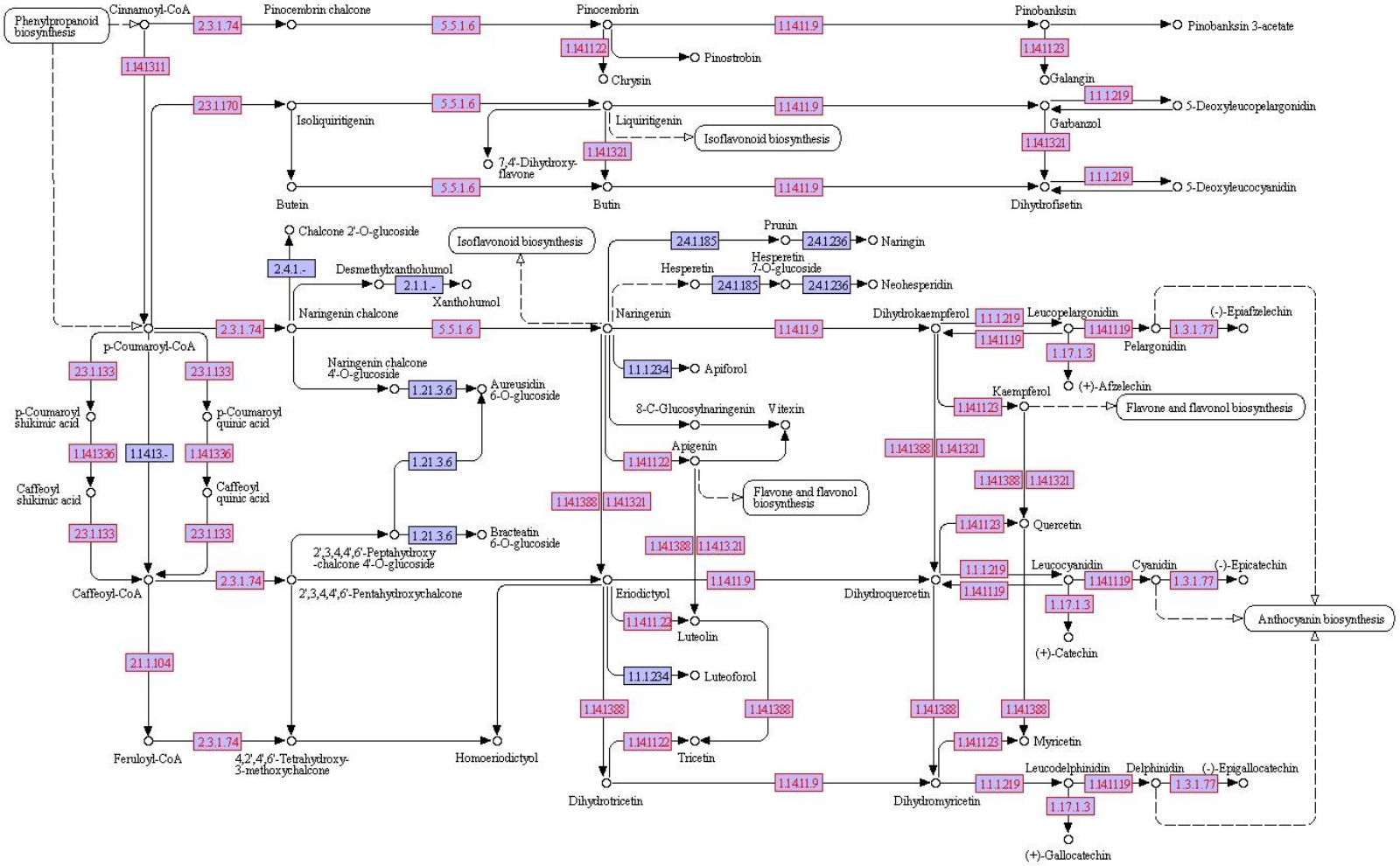

Figure 1.6 | Flavonoid biosynthetic pathway. The 19 EC numbers in red boxes were found in *P. lobata* dataset.

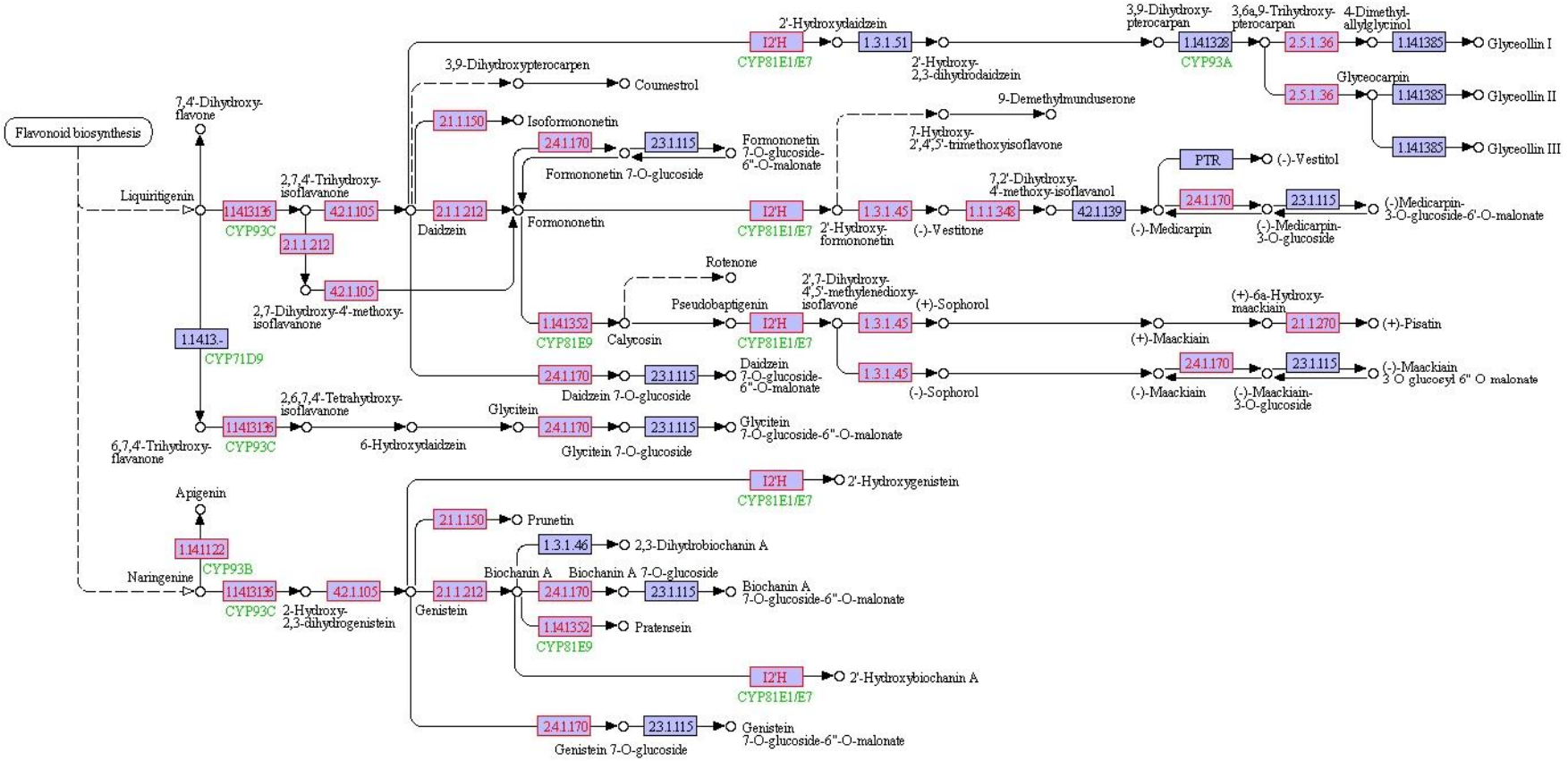

Figure 1.7 | Isoflavonoid biosynthetic pathway.

Fourteen enzyme commission numbers in red boxes were found in *P. lobata* dataset.

1.3.7 Genes Involved in Isoflavonoid Biosynthesis in *P. lobata*

Flavonoids are a group of polyphenolic compounds distributed widely throughout the plant kingdom. These compounds modulate the activity of enzymes to benefit the entire organism. As an important subgroup of flavonoids, isoflavonoids are mainly produced in legumes and affect oxidative stress markers, immune function and adipogenesis (Miadokova, 2009).

In the phenylpropanoid pathway, the synthesis of flavonoids is initialized by transforming phenylalanine into *p*-coumaroyl-CoA. To initiate flavonoid biosynthesis, chalcone synthase catalyzes the formation of chalcone scaffolds, from which all flavonoids derive (Falcone Ferreyra et al., 2012; Saito et al., 2013).

Based on our functional annotation findings, 45 contigs were predicted to represent seven enzymes critical to the biosynthesis of daidzein, which may be necessary to make the puerarin found in *P. lobata*. The number of contigs corresponding to each enzyme and the biosynthesis procedure are presented in **Figure 1.8**.

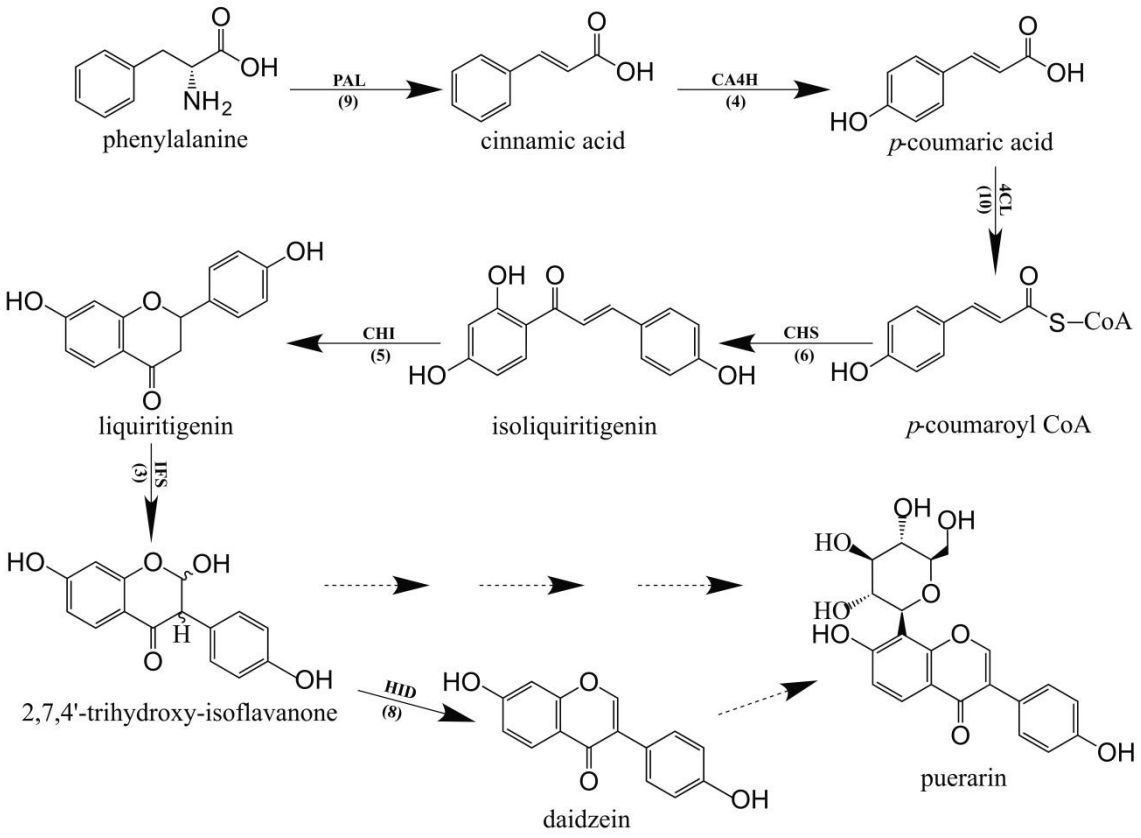

Figure 1.8 | Proposed daidzein biosynthesis pathway in *P. lobata*. Every enzyme is followed by the number of corresponding contigs in parentheses. PAL: Phenylalanine ammonia-lyase, EC 4.3.1.24; CA4H: *Trans*-cinnamate 4-monooxygenase, EC 1.14.13.11; 4CL: 4-coumarate-CoA ligase, EC 6.2.1.12; CHS: 6'-deoxychalcone synthase, EC 2.3.1.170; CHI: Chalcone isomerase, EC 5.5.1.6; IFS: 2-hydroxyisoflavanone synthase, EC 1.14.13.136; HID: 2-hydroxyisoflavanone dehydratase, EC 4.2.1.105. Solid arrows show the pathways identified by the data obtained while the dotted arrows show unsolved steps.

Figure 1.9 shows the expression profile for the 45 contigs corresponding to seven daidzein-biosynthesis-related enzymes across 5 libraries. From the heatmap, majority of the contigs had strong expression in library 4 (young root) and high correlation could be found among PAL, 4CL, CHS, CHI and HID, which gave light to the biosynthetic pathway of flavonoids in *P. lobata*.

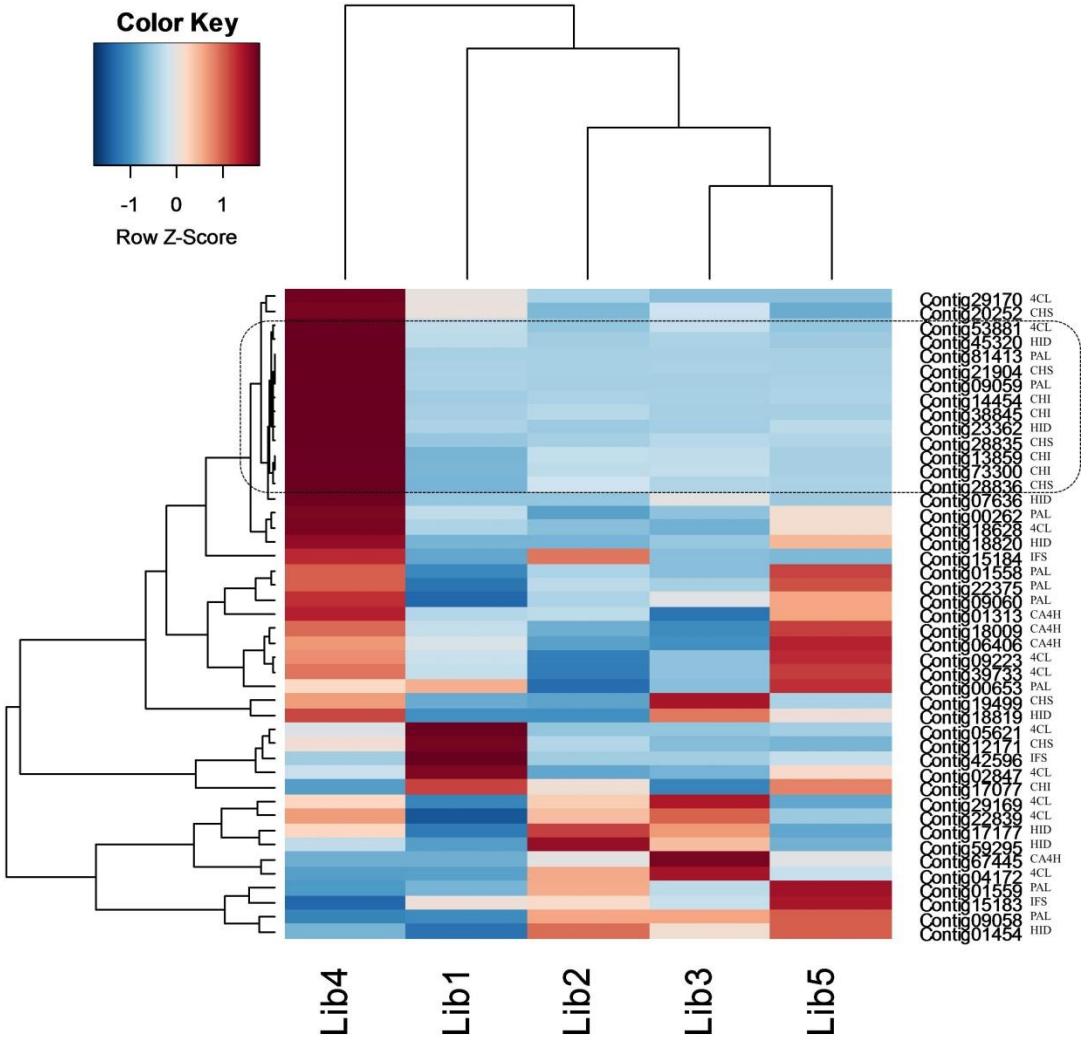

Figure 1.9 | Heatmap showing the expression profile for 45 contigs related to daidzein biosynthesis. Next to each contig name, the enzyme abbreviation is presented.

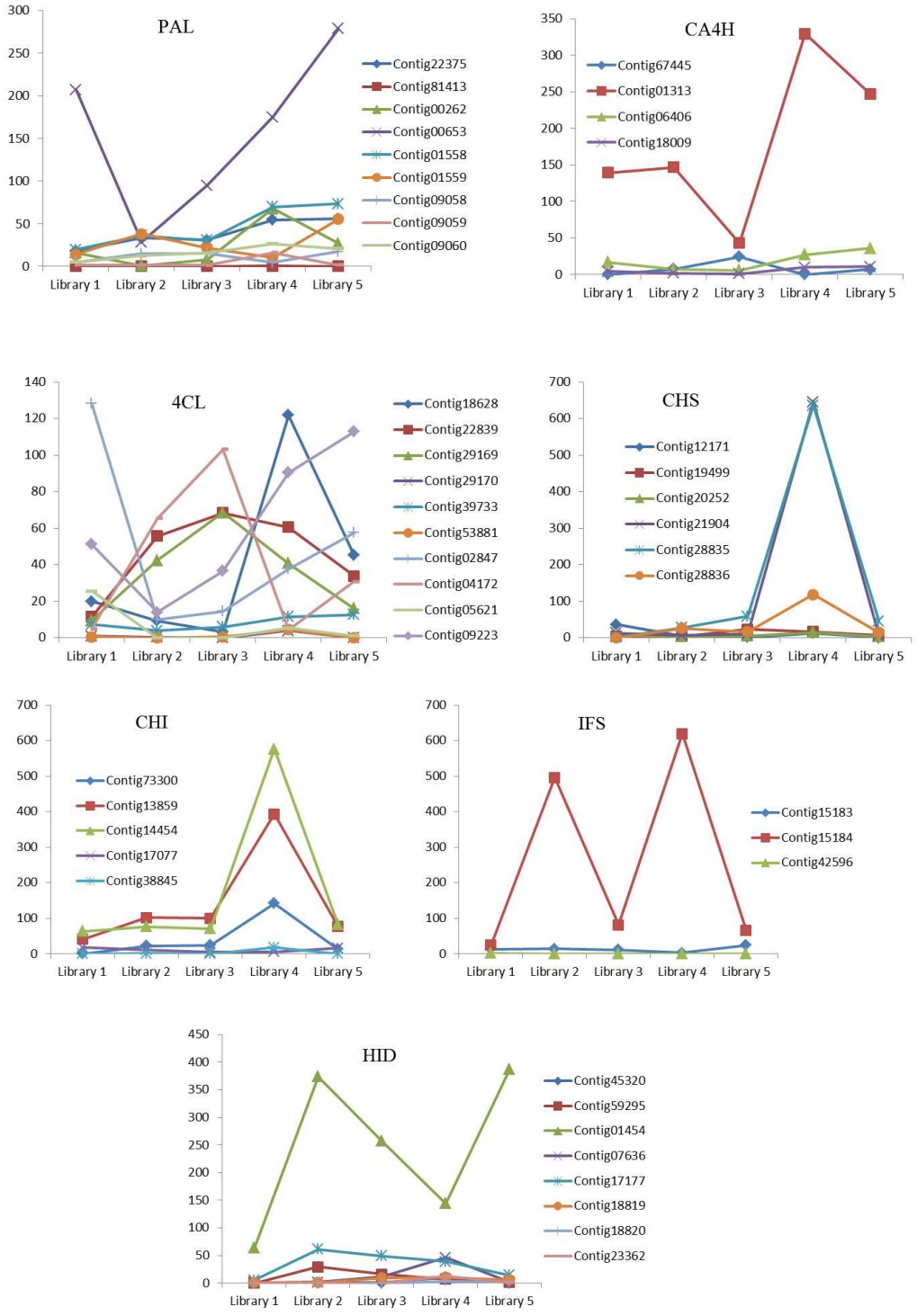

Figure 1.10 | Expression profile for isoflavonoid biosynthesis contigs. The vertical axis indicates RPKM values of the annotated contigs while the horizontal axis shows five libraries in *P. lobata* dataset.

A previous study (Chen et al., 2001) showed that concentrations of puerarin and daidzein were up to 3-fold higher in the roots as compared to the veins of *P. lobata*. In our study, biosynthesis of the chalcone scaffold showed that the expression of downstream enzymes exhibits a clear pattern that provides evidence into the organ-specific biosynthesis of daidzein (**Figure 1.10**). In *P. lobata*, early steps for the biosynthesis of daidzein and even puerarin might take place mainly in young root. After transportation of the required precursors to other parts of the plant or along with organ growth, the expression of related enzymes increases in mature root. Finally, the high expression of 2-hydroxyisoflavanone dehydratase in stem and mature root may account for the accumulation of daidzein and puerarin in certain plant organs.

Figure 1.8 demonstrates 45 contigs involved in the isoflavonoid biosynthetic pathway. We carried out to validate the genuine biological expression profile of such contigs, focusing on the genes from chalcone synthase (EC 2.3.1.170) to 2-hydroxyisoflavanone dehydratase (EC 4.2.1.105) since these are crucial factors leading to the accumulation of isoflavonoids. According to **Figure 1.10**, several contigs can be aligned to each of the four enzymes. Therefore, we set the following criteria for selecting the appropriate candidate contigs in qRT-PCR experiment: firstly, the contig itself should consist of long fragment; secondly, the expression judging by RPKM values across the five tissues was relatively high; thirdly, the identity between the candidate contig and target enzyme should be high. Contig 21904, 14454, 15184 and 01454, corresponding to chalcone

synthase, chalcone isomerase, 2-hydroxyisoflavanone synthase and 2-hydroxyisoflavanone dehydratase, respectively, were chosen to perform qRT-PCR. For CHS and CHI, **Table 1.3** shows similar pattern of gene expression profile between deep transcriptomic data and qRT-PCR result. Regarding chalcone synthase, its differential expression between young root and leaf was detected by NOISeq-sim and the ratio of $2^{-\Delta Ct}$ was also the highest one, 36. However, although the trend that the expression of 2-hydroxyisoflavanone synthase (IFS) and HID in library 1 was always lower compared to other libraries was verified by both approaches, the ratio varied.

Table 1.3 Validation of differentially expressed genes related to isoflavonoid biosynthesis

Contig	Annotation	Fold change : RNA-Seq and qRT-PCR validation							
		Lib 2 / Lib 1		Lib 4 / Lib 1		Lib 3 / Lib 1		Lib 5 / Lib 1	
		RPKM	$2^{-\Delta Ct}$	RPKM	$2^{-\Delta Ct}$	RPKM	$2^{-\Delta Ct}$	RPKM	$2^{-\Delta Ct}$
21904	CHS	0.49	1.87	48*	36	0.76	0.97	0.47	0.60
14454	CHI	1.20	0.72	9.0	14.6	1.12	0.63	1.29	2.07
15184	IFS	20.1	2.90	25.1	6.84	3.31	2.45	2.69	2.14
01454	HID	5.87	7.86	2.27	12.3	4.04	8.51	6.07	25.1

Fold change of RPKM and $2^{-\Delta Ct}$ between Lib 1 (Library 1) and other Libraries, respectively. *differently expressed gene detected by NOISeq-sim. CHS chalcone synthase, CHI Chalcone isomerase, IFS 2-hydroxyisoflavanone synthase, HID 2-hydroxyisoflavanone dehydratase.

For example, RPKM value suggested 20.1-fold elevated expression in library 2 while qRT-PCR granted only 2.90-fold. We collected the samples for sequencing in 2012 but the materials used for qRT-PCR were obtained in 2014 from the same location, though. The slightly changed sampling conditions may result in the variations in validation experiment. The primers designed for qRT-PCR are listed in **Table 1.4**.

Table 1.4 Primers designed for qRT-PCR experiment

Contig	Annotation	Forward primer	Reverse primer
00518	β-actin	TCCACTGGCATACAGAGACAAGA	GGCACCACTCAATCCCAAG
21904	CHS	AATGGCTGCCACCTTAGTCTCT	TCTTTTGTGGTAACTGTGCTGGTT
14454	CHI	GCAGTTTTCCATCACCTTCTTTG	GCTGGTTGAGACCCTTGACTTCT
15184	IFS	CTGTTGGGCCTCTGCACTTT	GTTCCCTTCGGACCTTACTGG
01454	HID	GCTTCCCACGCCAACAA	CCGCTGGTTTCACCTCCTAC

CHS, chalcone synthase; CHI, chalcone isomerase; IFS, Isoflavanone synthase; HID, 2-hydroxyisoflavanone dehydratase.

1.4 Discussion

From the expression pattern of isoflavonoid biosynthesis in Kudzu, transportation of certain precursors into other parts of the plant for downstream reaction may be required (refer to genes involved in isoflavonoid biosynthesis in *P. lobata*). By searching the annotation data, 41 expressed ABC transporters were retrieved. Regarding IFS and HID, suggested by the changes of expression level in different tissues, transportation of the intermediates may occur.

Table 1.5 Contigs annotated as ABC transporter showing Pearson correlation coefficients to contig 01454 and contig 15184.

Contig_ID	Coeffiecient_to _contig01454	Coefficient_to _contig_15184	Annotation
Contig00240	-0.73	-0.18	ABC transporter D family member
Contig00241	-0.77	-0.5	ABC transporter D family member
Contig02360	0.34	-0.21	ABC transporter family
Contig02621	-0.56	-0.72	ABC transporter family protein
Contig03204	0.2	-0.23	ABC transporter family
Contig07356	-0.63	0.6	ABC transporter G family member
Contig07381	-0.78	-0.39	ABC transporter family protein
Contig09681	-0.78	-0.41	AAA ATPase; ABC transporter, transmembrane region, type 1
Contig09727	0.05	0.58	ABC transporter C family member
Contig10273	-0.38	0.69	ABC transporter G family member
Contig11124	-0.79	-0.44	ABC transporter I family member
Contig16119	-0.4	0.73	ABC transporter family protein
Contig16951	-0.97	-0.24	ABC transporter D family member
Contig17808	-0.96	-0.28	ABC transporter D family member
Contig18533	-0.91	0.06	ABC transporter C family member

Contig22029	-0.22	0.6	ABC transporter C family member
Contig23310	-0.72	-0.47	ABC transporter B family member
Contig27958	-0.3	-0.58	ABC transporter C family member
Contig27959	-0.38	0.58	ABC transporter C family member
Contig31585	-0.92	-0.11	ABC transporter C family member
Contig33504	0.05	0.37	Multidrug resistance protein ABC transporter family
Contig33505	0.51	0.06	Multidrug resistance protein ABC transporter family
Contig33602	-0.65	-0.5	ABC transporter D family member
Contig36545	0.08	0.2	ABC transporter B family member
Contig41300	0.5	0.22	ABC transporter G family member
Contig41301	0.28	-0.17	ABC transporter G family member
Contig42074	-0.06	-0.5	ABC transporter G family member
Contig44359	-0.08	-0.62	ABC transporter B family member
Contig56331	-0.72	-0.47	ABC transporter ATP-binding protein/permease
Contig58577	-0.58	-0.79	ABC transporter B family member
Contig69143	-0.4	0.73	ABC transporter ATP-binding protein/permease
Contig69144	-0.54	-0.67	ABC transporter ATP-binding protein/permease
Contig70145	-0.4	0.73	ABC transporter ATP-binding protein/permease
Contig73108	0.06	0.36	Multidrug resistance protein ABC transporter family
Contig73572	0.43	-0.12	Multidrug resistance protein ABC transporter family (Fragment)
Contig74670	-0.34	0.69	ABC transporter G family member
Contig76477	0.26	0.89	Multidrug resistance protein ABC transporter family (Fragment)
Contig76628	0.43	-0.12	Multidrug resistance protein ABC transporter family
Contig79472	-0.4	0.73	ABC transporter B family member
Contig80235	-0.4	0.73	ABC transporter B family member
Contig81715	0.02	1	ABC transporter B family member

Table 1.5 lists 41 ABC transporters found in Kudzu dataset along with Pearson correlation coefficients to contig 15184 which is a putative IFS and contig 01454, an annotated HID.

A recent study provided fairly constructive insights into the biosynthetic pathway of puerarin and contributed more than 6,365 ESTs (He et al., 2011). We integrated the publically available ESTs into our raw reads from five different tissues of Kudzu and then performed *de novo* assembly altogether. This enabled us to utilize related information and the assembled contigs showed identical or highly similar transcripts with the ESTs regarding glucosyltransferase.

Many *C*-glucosyltransferases have been identified in bacteria, insects and plants, especially in cereals. The elucidated mechanism for *C*-glycosylation of flavonoids proved 2-hydroxylation of flavanones was the appropriate premise for the catalytic reaction to proceed. Likewise, in studying the biosynthesis of puerarin, 2-hydroxylation of isoflavanone (2,7,4'-trihydroxy-isoflavanone) should be considered as a possible substrate for its formation when daidzein as a direct putative precursor meets with obstacles. With the formation of trihydroxy-isoflavone 8-*C* glycoside catalyzed by suitable UDP-dependent glucosyltransferases, the glycoside may be subjected to dehydration reaction, resulting in puerarin. 49 contigs were annotated as glucosyltransferase in our dataset, if the target glucosyltransferase utilizes either one of the above-mentioned precursors, the correlation with the enzyme directly producing 2,7,4'-trihydroxy-isoflavanone or daidzein would be significant. **Table 1.6**

lists the annotated glucosyltransferases along with Pearson correlation coefficients to HID and IFS.

Table 1.6 Putative glucosyltransferases with Pearson correlation coefficients to HID and IFS

Contig_ID	Coefficient_to_HID	Coefficient_to_IFS	Annotation
Contig02990	0.37	0	Sucrose-UDP glucosyltransferase
Contig03131	-0.4	0.74	Anthocyanidin 3-O-glucosyltransferase
Contig03133	-0.45	0.48	Anthocyanidin 3-O-glucosyltransferase
Contig05593	-0.59	0.53	glycoprotein glucosyltransferase
Contig09646	-0.68	0.31	Cytokinin-O-glucosyltransferase
Contig10691	-0.59	0.45	glycoprotein glucosyltransferase
Contig11620	0.79	0.1	Putative glucosyltransferase
Contig11621	-0.69	-0.4	Putative glucosyltransferase
Contig11622	-0.84	-0.39	Putative glucosyltransferase
Contig12257	-0.66	-0.54	Cytokinin-O-glucosyltransferase
Contig14425	-0.77	0.47	Putative UDP-glucosyltransferase
Contig15137	-0.6	0.54	glycoprotein glucosyltransferase
Contig15603	-0.72	-0.47	Isoflavonoid glucosyltransferase
Contig20530	-0.78	-0.3	Sterol 3-beta-glucosyltransferase
Contig22923	-0.07	0.2	Sterol 3-beta-glucosyltransferase
Contig22924	-0.94	0.04	Sterol 3-beta-glucosyltransferase
Contig23483	-0.4	0.73	Zeatin O-glucosyltransferase
Contig23956	-0.35	0.79	Isoflavonoid glucosyltransferase
Contig25035	-0.19	0.94	Putative glucosyltransferase
Contig28030	-0.29	-0.39	flavonoid 3-O-glucosyltransferase
Contig28462	-0.03	0.46	Hydroquinone glucosyltransferase
Contig29838	-0.47	0.22	Isoflavone 7-O-glucosyltransferase 1
Contig29839	0.15	-0.43	Isoflavone 7-O-glucosyltransferase 1
Contig31158	0.34	-0.12	Cytokinin-O-glucosyltransferase
Contig32270	-0.39	0.71	Anthocyanidin 3-O-glucosyltransferase
Contig32277	-0.13	0.58	Hydroquinone glucosyltransferase
Contig40732	-0.67	0.07	Sucrose-UDP glucosyltransferase

Contig42441	-0.9	-0.18	Cytokinin-O-glucosyltransferase
Contig43033	0.27	-0.7	Isoflavone 7-O-glucosyltransferase 1
Contig44304	-0.15	0.59	Isoflavone 7-O-glucosyltransferase 1
Contig45936	-0.04	-0.98	Isoflavone 7-O-glucosyltransferase 1
Contig46490	0.34	-0.55	Anthocyanidin 3-O-glucosyltransferase
Contig47890	-0.28	0.75	Sterol 3-beta-glucosyltransferase
Contig51492	-0.41	0.7	UDP-glucosyltransferase
Contig52569	-0.22	0.64	Sterol 3-beta-glucosyltransferase
Contig65011	-0.4	0.68	UDP-glucosyltransferase
Contig73434	0.56	-0.38	Limonoid UDP-glucosyltransferase
Contig74493	0.1	0.21	Anthocyanidin 3-O-glucosyltransferase
Contig77022	-0.35	0.47	Anthocyanidin 3-O-glucosyltransferase
Contig79587	-0.4	0.73	Cytokinin-O-glucosyltransferase
Contig79864	-0.4	0.73	Cytokinin-O-glucosyltransferase
Contig03787	0.47	-0.6	Glucosyltransferase
Contig06374	-0.83	0.07	Glucosyltransferase-13 (Fragment)
Contig11423	-0.71	-0.49	Glucosyltransferase
Contig14082	-0.58	-0.42	Glucosyltransferase-2
Contig14083	-0.6	-0.35	Glucosyltransferase-12
Contig14085	0.3	0.89	Glucosyltransferase-2
Contig24631	0.04	0.67	Glucosyltransferase-5
Contig28995	-0.33	0.73	Glucosyltransferase-12

Kudzu root, which is the main part prescribed in oriental medicines in treating various diseases, produce predominantly isoflavone *C*- and *O*-glucosides. We collected five tissues from which the deep transcriptomic data were generated and studied puerarin and daidzin profile using HPLC. Both puerarin and daidzin are highly accumulated in mature root and root vascular cylinder and the concentration of puerarin is higher than that of daidzin (**Table 1.7**). Although several genes related to isoflavonoid

biosynthesis are highly expressed in young root, the concentration of the two compounds is low in young root compared with that in mature root. This may be due to the essential enzymes for the production of puerarin actively expressed in the young and mature roots but the accumulation of puerarin does not reach to the maximum yet in the young root. Deep transcriptomic data obtained in this study may provide the key to this question. In relatively young stem which was used in this study and leaf, daidzin was not detectable.

RNA-Seq analysis is cost effective and the most efficient approach currently available to manage high-throughput data. By consolidating data information obtained from five *P. lobata* libraries, we analyzed the differential expression profile and mapped biosynthetic pathways against KEGG using enzyme accession numbers. Evaluating overrepresented GO terms by considering the RPKM values of the corresponding contigs provided a more accurate representation of the data. By qRT-PCR and HPLC, both gene expression validation and metabolite analysis were performed. The deep transcriptomic data we present here may facilitate future research on this promising plant.

Table 1.7 Determination of puerarin and daidzin in fresh plant samples

	Leaf	Stem	Mature root	Young root	Root VC
Puerarin	N.D.[a]	1.473± 0.007 [b]	4.024 ± 0.005	0.231 ± 0.002	3.327 ± 0.005
Daidzin	N.D.	N.D.	0.668 ± 0.004	0.156 ± 0.002	0.994 ± 0.008

[a]N.D. = Not detected.

[b]Means ± SD (mg/g fresh weight, n = 3).

Chapter 2: Transcriptome Analysis of Nine Tissues to Discover Genes Involved in the Biosynthesis of Active Ingredients in *Sophora flavescens*

2.1 Introduction

In 1889, the characteristic compound matrine was isolated from the dry roots of *S. flavescens* (Nagai, 1889) and several decades later, the absolute structure of (+)-matrine was figured out (Okuda et al., 1966). Due to its notable medicinal efficacy, attempts to synthesize and biosynthesize matrine were conducted (Boiteau et al., 1998; Saito et al., 1989; Shibata and Sankawa, 1963). In 1995, Saito et al. proposed the biosynthetic pathway of the carbon framework of matrine (Saito and Murakoshi, 1995). Although the first steps of quinolizidine alkaloids biosynthesis have been elucidated recently in *S. flavescens* (Bunsupa et al., 2012), more effort needs to be done to puzzle out the practical biosynthetic pathway of matrine.

S. flavescens also contain series of flavonoids and isoflavonoids such as kuraridin, kurarinone, isokurarinine, daidzein, maackiain (Chen et al., 2004). Much effort has been done to elucidate the biosynthetic pathway of flavonoids and many of the related genes are clear now in other species (Falcone et al., 2012), but the ones involved in the biosynthesis of flavonoids in *S. flavescens* have not yet been discovered. In spite of several research findings cracking relevant quinolizidine alkaloids biosynthetic

pathway and membrane-bound prenyltransferase (Gao et al., 2011; Sasaki et al., 2011), to our best knowledge, analysis using next generation sequencing approach for *S. flavescens* was not found up to date.

More and more genomes of model organisms have been sequenced (Meyer et al., 2013; Michael and Jackson, 2013). Nevertheless, for those non-model plants, lack of reference genome information jeopardizes studies on the underlying genes which are involved in biological processes related to vital plant physiology and drug development, etc. With this regard, transcriptome sequencing plays an essential role in apprehending the genetic diversity of organisms. Furthermore, such approaches help to obtain overall insights into whole gene sets associated to the protein diversity (Saito, 2013). Armed with well-established approaches such as Short Oligonucleotide Analysis Package (SOAPdenovo) (Li et al., 2009), Assembly by Short Sequences (AbySS) (Simpson et al., 2009), Trinity (Grabherr et al., 2011), transcriptome profiling accelerates its pace in processing tremendous amount of data generated from large-scale sequencing projects. Despite the intensive challenges including library construction, reducing errors in image analysis and removal of low-quality reads, massively parallel cDNA sequencing (RNA-Seq) offers a more precise measurement of levels of transcripts and their isoforms than other methods (Wang et al., 2009).

In this study, 203,598,590 fastq format reads from 9 tissues of *S. flavescens* were generated by Illumina's next-generation sequencing approach. CLC Genomics Workbench (CLC bio, Denmark) was

subsequently applied to conduct *de novo* assembly. Based on the findings provided by Gene Ontology and KEGG pathway mapping, the candidate genes that may be involved in the biosynthesis of key chemical components were identified. By studying expression pattern of genes related to the biosynthesis of quinolizidine alkaloids, we propose some promising contigs for future research.

2.2 Materials and Methods

2.2.1 Sampling and Total RNA Extraction

Except for leaf sample collected in May 2012, all the rest fresh tissues and organs were obtained from healthy *S. flavescens* plants growing in Chiba, Japan in June 2013. Callus tissue whose origin and subculturing were described previously (Yamamoto et al., 1991), and other eight parts of the plant were sampled, including leaf, flower, stem, young bud, mature bud, bud right before blossom (BBB), pedicel while bud stage (PBS), pedicel while blossom (PWB) (**Figure 2.1**).

After sampling from the field, the tissues were dipped into RNA stabilization solution (RNA*later*, Life technologies, USA) immediately. Then RNA*later* solution was gently removed with a Kimwipe and the remaining sample was frozen by liquid nitrogen and subsequently powdered by using Multi Beads Shocker (Yasui Kikai, Japan).

Total RNA was extracted from powdered tissues of *S. flavescens* by TRIzol Reagent (Invitrogen, USA) according to the instructions, respectively. The acquired RNA was then cleaned up by RNeasy Mini Kit

(Qiagen, USA).

For semi-quantitative RT-PCR analysis, in November 2014, the tissues of root, stem and leaf of kurara were sampled in the same location, where the other tissues were collected for deep-transcriptome sequencing, and subjected to the extraction of total RNA as described above.

Figure 2.1 | Different organ / tissue of *S. flavescens* used for total RNA extraction. 1 callus, 2 leaf, 3 flower, 4 stem, 5 young bud, 6 mature bud, 7 bud right before blossom (BBB), 8 pedicel while bud stage (PBS), 9 pedicel while blossom (PWB).

2.2.2 cDNA Library Construction and Illumina Sequencing

TruSeq RNA Sample Prep Kit v2 (Illumina, USA) was applied according to

the manufacturer's recommendation. First of all, the mRNA portion in total RNA was polyA-selected and subsequently fragmented. Then the double-stranded cDNA was prepared for cDNA library construction. After the blunt-end fragments were generated, followed by indexed adaptor ligation, the samples were then hybridized to flow cells. Cluster amplification was performed using the cBot Cluster Generation System (Illumina, USA) and finally sequenced by Illumina's next-generation sequencing instrument.

2.2.3 D*e novo* Assembly adopting CLC Genomics Workbench

Before *de novo* assembly, the original fastq format data of *S. flavescens* were subjected to CLC trimming process for the purpose of eliminating empty reads, reads with poor quality and as a result, clean reads were obtained. The CLC workbench (version 6.5) was then utilized to process the clean reads and all contigs over 300 bp were taken into consideration for downstream work.

Since the assembled contigs may contian duplicates due to sequencing error, CD-HIT-EST was utilized with representative sequences at 90% identity to obtain unique unigenes.

2.2.4 RPKM Calculation and Expression Analysis

The standard formula for RPKM (reads per kilobase of the transcript per million mapped reads) is as follows: $\text{RPKM} = \frac{10^9 \times C}{N \times L}$ (Chen et al., 2011). In order to estimate the expression level of the contigs, with the aid of

Burrows-Wheeler Aligner (Li and Durbin, 2009), Sequence Alignment/ Map tools (Li et al., 2009) and High Throughput Sequencing (HTSeq) (Anders, 2010), we applied the Mortazavi's approach to map all the fastq format reads back to the contigs and calculated the RPKM values (Mortazavi, et al., 2008). Because the *S. flavescens* samples lacked biological replicates, NOISeq-sim, the well-established non-parametric approach for the identification of differentially expressed (DE) genes, was used to analyze 36 independent pair-wise sample comparisons (Tarazona et al., 2011).

2.2.5 Annotation Pipeline and Data Mining

After applying CD-HIT-EST, All 83,325 *de novo* contigs were subjected to BLASTx against the non-redundant (NR) protein database at NCBI and the Universal Protein resource (UniProt) at UniProt consortium with the e-value threshold set at 1e−10.

If possible, the top 20 subject sequences from databases to show alignment would be taken into consideration. As to the large BLASTx result, only percent identities over 40% and e-values less than 1e−30 were utilized for downstream analysis. After eliminating redundancies, online ID mapping was conducted by uploading all unique gene identifiers in fasta format to the uniprot official website (http://www.uniprot.org) to retrieve accession information.

By combining the returned target list and the UniProt accession numbers (ACs), we finished annotation process using the same online

facilities. Reviewed findings from UniProtKB/Swiss-Prot as well as UniProtKB/TrEMBL were studied for data mining. In all, 1,350 enzyme commission (EC) numbers were applied to map pathways against KEGG, and the enzymes related to flavonoids and isoflavonoids biosynthesis were studied.

2.2.6 Semi-quantitative Reverse Transcription PCR for a putative lysine/ ornithine decarboxylase (*L/ODC*) gene

By using the total RNA extracted from the tissues harvested in 2014, cDNA was prepared according to the manufacturer's instructions, SuperScript VILO kit (Invitrogen, USA).

A 477 bp fragment of *L/ODC* was amplified by PCR using Ex *taq* DNA polymerase (Takara) and specific primers (L/ODC-F: 5'-GAC ATT GGT GGC GGT TTC AC-3', L/ODC-R: 5'-AGT GCT AAA GCC ATT GAA GTT GG-3'). The PCR for *L/ODC* cDNA was performed with an initial denaturation at 94℃ for 2 min, then 26, 28 or 30 cycles each at 94℃ for 30 s, at 54℃ for 30 s, and 72℃ for 50s.

For normalization of the different RNA preparations, a 571 bp fragment of *S. flavescens β-actin* was amplified with the following primers: (Act-F: 5'-AAG GCC AAC AGA GAG AAG ATG AC-3', Act-R: 5'-ACC CAC CAC TAA GCA CGA TAT TT-3'). The PCR for *β-actin* cDNA was performed with an initial denaturation at 94℃ for 2 min, then 22 cycles each at 94℃ for 30 s, at 53℃ for 30 s, and 72℃ for 50s.

The PCR products were separated by electrophoresis using 1.5% gel at 100V, and the gel was further stained by Ethidium Bromide Solution (Nacalai Tesque Inc., Japan).

2.2.7 *S. flavescens* CYP86A24-like full-length cDNA amplification

Total RNA was prepared from relatively young leaves of *S. flavescens* and then SuperScript Ⅱ first-strand cDNA synthesis (Invitrogen, USA) was applied according to manufacturer's instructions. To amplify the full-length sequence and be ready for the subsequent experiment using Gateway system (Invitrogen, USA), we used the following primer pairs: (SfCYP-F: 5'-AAA AAG CAG GCT TCA CCA TGG ATG GAT GCA TCA ACG GCT TTT ATGA-3', SfCYP-R: 5'-AGA AAG CTG GGT CTC AAG CAT CAG CAG CAA CCA TTTC-3'). attB-PCR product was separated by electrophoresis and the target band was excised and purified according to PEG purification protocol (Invitrogen, USA).

2.2.8 Characterization of *S. flavescens* CYP86A24-like enzyme using Gateway system

An entry clone was generated by performing BP reaction using the purified attB-flanked DNA fragment and *pDONR* 221 vector (Invitrogen, USA), followed by LR reaction to create an expression clone with pYES-DEST52 and subsequent expression in *Saccharomyces cerevisiae*.

2.3 Results

2.3.1 *S. flavescens* Total RNA Preparation

The principal bioactive constituents of *S. flavescens* are the major quinolizidine alkaloids matrine and oxymatrine (Ling, et al., 2007). The contents of these two components in *S. flavescens* are also the main valuation criteria of this plant. Meanwhile, considerable quantity of flavonoids and isoflavonoids are also accumulated in *S. flavescens* (Shen, et al., 2013).

In this study, we aimed to collect information about the nature of the genes responsible for the biosynthesis of matrine and oxymatrine, and study the related genes involved in the biosynthesis of flavonoids and isoflavonoids in *S. flavescens*.

We extracted total RNA from the 9 tissues of this plant, resulting in 9 distinct cDNA libraries. We will refer to the libraries in the following manner: Lib 1 (callus), Lib 2 (leaf), Lib 3 (flower), Lib 4 (stem), Lib 5 (young bud), Lib 6 (mature bud), Lib 7 (bud right before blossom), Lib 8 (pedicel while bud stage), and Lib 9 (pedicel while blossom).

2.3.2 Next generation Sequencing and CLC *de novo* Assembly

Using the Illumina HiSeq platform, fastq format paired-end reads were generated for all 9 libraries of *S. flavescens*. Raw reads with poor quality were trimmed using CLC trimming function. Nucleotide fastq sequences of Illumina trimmed reads were deposited at DDBJ Sequence Read Archive (DRA) with the accession number DRA003182.

To practice *de novo* assembly, the fastq reads from all 9 libraries were combined together to generate fasta format contigs. After the entire run was assembled, the quality of the assembly (e.g., the ratio of aligned reads, average contig size, N75, N50 and N25 contig length) was identified to be better than the *de novo* results from individual libraries. The assembly of the 9 libraries was utilized in the following analysis to discover their genetic information.

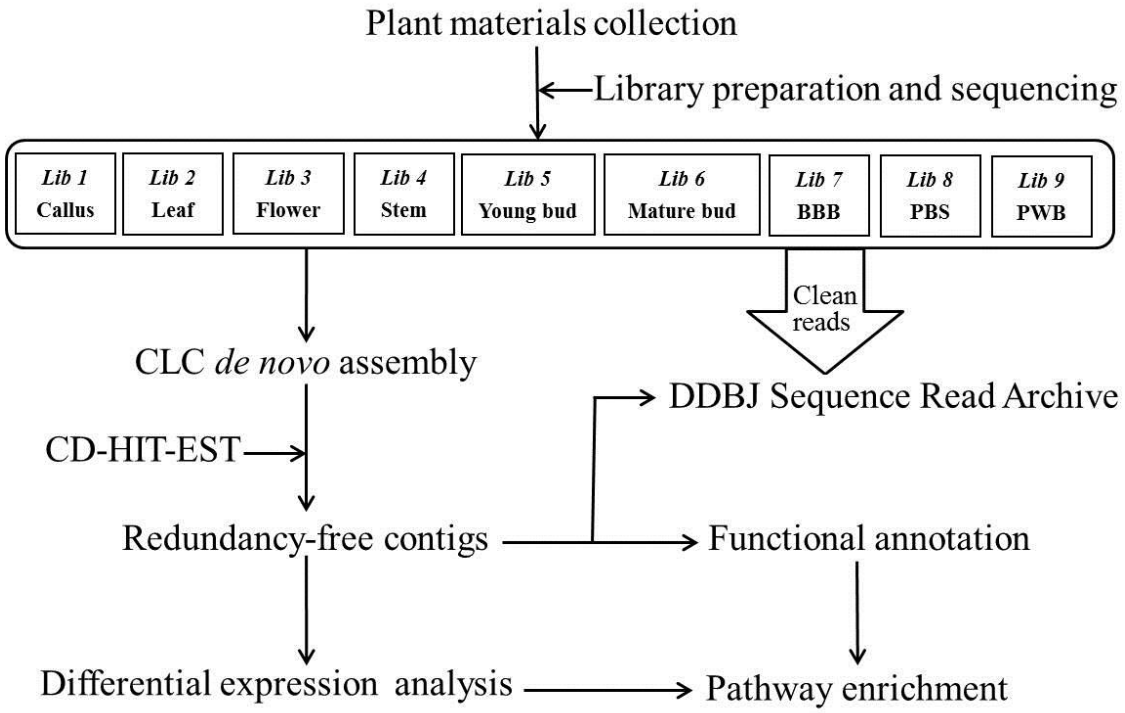

Figure 2.2 | Summary of the experimental design and analysis pipeline.

BBB, bud right before blossom; PBS, pedicel while bud stage;

PWB, pedicel while blossom.

85,054 contigs were generated using CLC genomics workbench. By applying CD-HIT-EST with a threshold set at 0.9, duplicate contigs were identified and discarded, resulting in 83,325 non-redundant contigs. An overview of the experimental pipeline is shown in **Figure 2.2**. **Table 2.1** summarizes sequencing and assembly results.

Table 2.1 Overview of *S. flavescens* transcriptome assembly

Items	Numbers
Total bases	20,044,281,186
Average length of reads (bp)	98.4
No. of reads	203,598,590
Average length of contigs (bp)	664
Maximum contig length (bp)	15,827
N75; N50; N25 (bp)	431; 969; 1,947
No. of contigs over 300 bp	85,054
Non-redundant contigs	83,325

2.3.3 *S. flavescens* GC Content Profile

The average GC content of *S. flavescens* transcripts was found to be 39.3% (**Figure 2.3**). Before 5000 base position, the GC content is relatively stable at the level of 40%. However, after 5000 base position, the GC content of *S. flavescens* varies dramatically due to the rapidly reduced number of the assembled contigs. The GC value found in *S. flavescens* is close to the reported legume plants *Glycine max* (43%) and *Medicago truncatula* (40%) (Kawaguchi and Bailey-Serres, 2005).

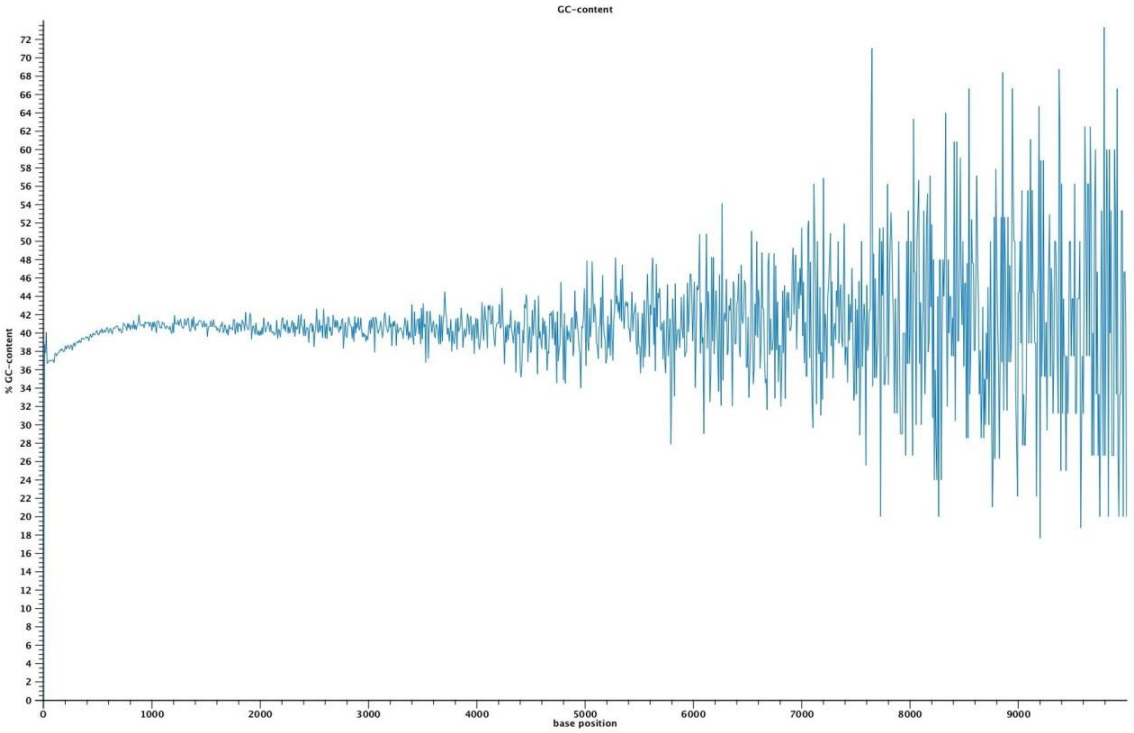

Figure 2.3 | GC content at different base positions for all *S. flavescens* contigs. The length of contigs varies from 300 to 15,827 bp. The GC content presented here does not include the information of contigs over 10,000 bp. Percentage of GC content stands for number of G- and C-bases observed at current position normalized to the total number of bases observed at that position.

2.3.4 NOISeq-sim Analysis on Differentially Expressed Transcripts

In order to perform transcript expression analysis, RPKM calculation is the initial step to do. RPKM value of the contigs allows such normalized output to be applied directly for the comparison of gene expression.

For DE analysis, 36 pair-wise comparisons of the nine libraries were conducted by applying NOISeq-sim, a non-parametric approach.

By running NOISeq-sim on the R language platform with a given threshold (q = 0.9) for selecting differentially expressed features, the resultant number of DE transcripts varied across comparisons. The highest value obtained was 1,061 differences between callus and mature bud transcripts; the lowest value obtained was 0 between mature bud and BBB (**Figure 2.4 and Figure 2.5**).

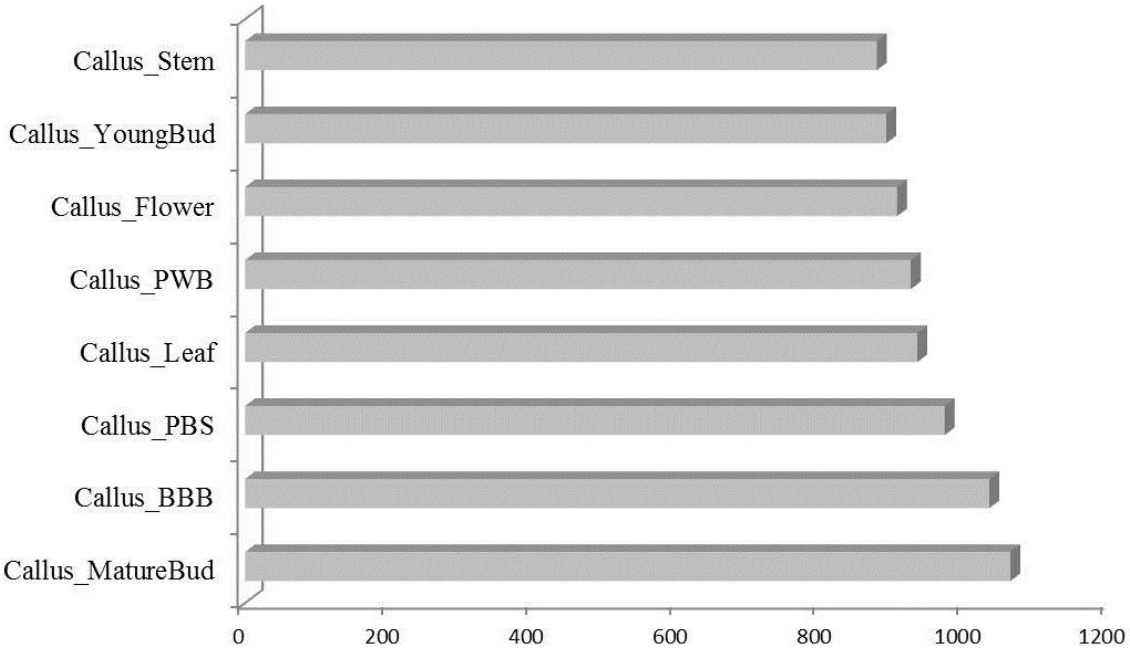

Figure 2.4 | Number of DE genes between callus and the rest 8 tissues.

The pair-wise comparisons were between callus and the rest 8 tissues, respectively, of *S. flavescens* with the number varying from 876 to 1,061.

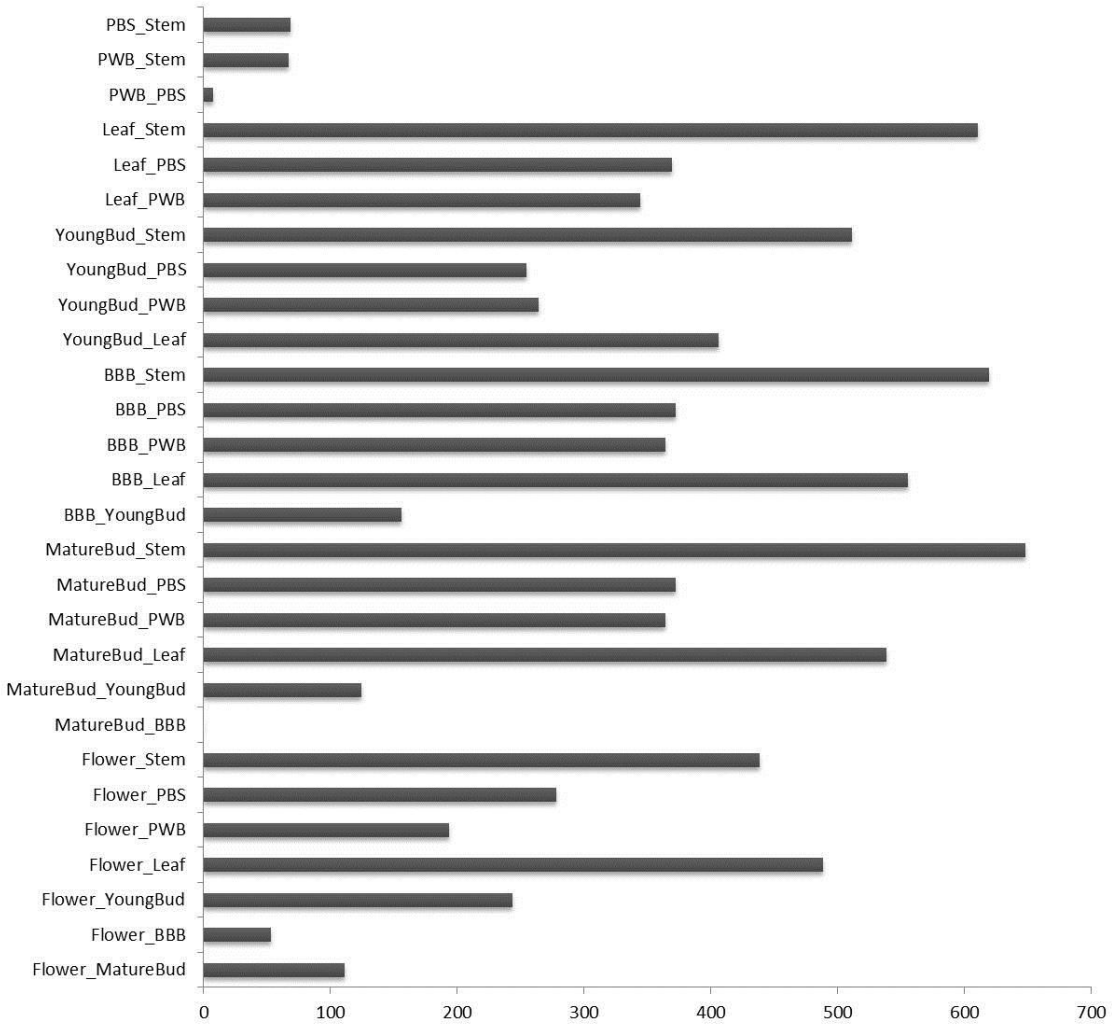

Figure 2.5 | Number of DE genes detected by NOISeq-sim.

The lowest DE genes is 0 between mature bud and BBB.

2.3.5 Transcripts Function Annotation and Gene Ontology Analysis

BLAST was applied to search protein databases to identify similar subject sequences. When the threshold E-value was set to 1e−10 and if possible, the top 20 subject sequences for each query sequence were taken into consideration, we obtained 570,611 subject sequences for all 83,325 query

sequences. In order to reduce redundancy and simplify the experiment procedure, more stringent criteria for retrieving the candidate genes were set. With this approach, significant matches were assigned to 27,909 transcripts.

Gene Ontology is a useful tool to study the character of annotated genes. By utilizing the BLAST results as well as Web Gene Ontology Annotation Plot software, 25,921 contigs yielded corresponding GO terms that could be further classified into 50 sub-categories: 14 related to cellular components, 13 to molecular function and 23 to biological processes (**Figure 2.6**).

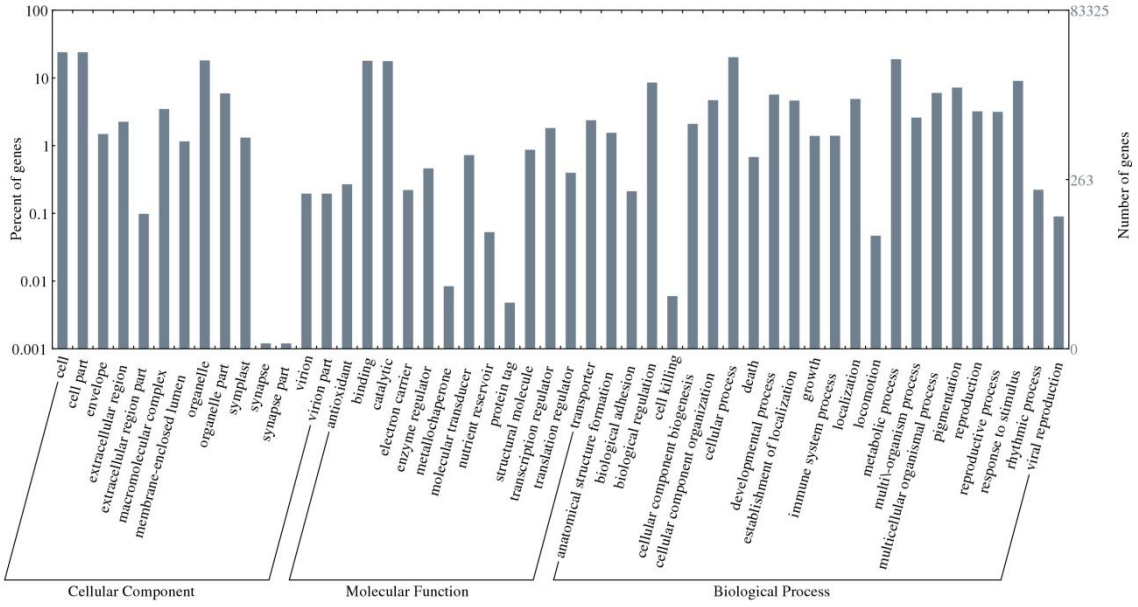

Figure 2.6 | Gene Ontology annotation for *S. flavescens* contigs.

Fifty sub-categories are affiliated to three main domains: molecular function, cellular components and biological processes.

Directly counting the number of GO terms related to the corresponding contigs bears a very obvious shortcoming. The expression level of those query sequences varies, which grants distinct weight to the same GO term as it corresponds to different query sequences. Regarding this, overrepresented GO terms were identified by Fisher's exact test. The one-tailed Fisher's exact *p*-values corresponding to overrepresented categories were calculated according to the counts in 2×2 contingency tables. Counts n_{11}, n_{12}, n_{21} and n_{22} in each table stand for: n_{11}, number of observations of a specific category in the first gene set; n_{12}, number of other categories in the first gene set; n_{21}, number of observations of a category in the second gene set; and n_{22}, number of observations of other categories in the second gene set (Takahashi et al., 2011).

For each *S. flavescens* library, contigs with RPKM value over 30.0 (the top ~10% of all transcripts) were regarded as highly expressed genes and extracted respectively. Then the merged 9,867 contigs were used to perform Fisher's exact test.

The overrepresented GO terms (**Supplementary 2.1**, GO terms with $p < 1E{-}30$ are listed) indicate expectedly cellular component like plasma membrane (GO: 0005886) and cytoplasm (GO: 0005737) plays crucial roles in all aspects regarding *S. flavescens*. Meanwhile, biological process including isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway (GO: 0019288) is also highlighted in the list, which will provide information for future research.

2.3.6 KEGG Pathway Mapping

ID mapping resulted in 1,350 unique enzymes from *S. flavescens* dataset and KEGG pathway mapping was performed. These enzyme commission numbers were assigned to 154 biological pathways with the largest number of enzymes (707) involved in metabolic pathways. Given the remarkable reputation of *S. flavescens* to accumulate large quantity of functional flavonoids, 18 flavonoid biosynthetic and 13 isoflavonoid biosynthetic enzymes are presented in **Figure 2.7** and **Figure 2.8**.

2.3.7 Putative Genes Involved in Isoflavonoid and Quinolizidine Alkaloids Biosynthesis in *S. flavescens*

From the essential amino acid phenylalanine, with the help of phenylalanine ammonia-lyase, cinnamic acid is generated. Then trans-cinnamate 4-monooxygenase catalyzes the reaction to produce 4-coumaric acid. 4-coumarate-CoA ligase produces the compound 4-coumaroyl CoA and now it is ready for the production of the specific flavonoids.

First, chalcone synthase plays the role to produce isoliquiritigenin, followed by the formation of liquiritigenin catalyzed by chalcone isomerase. Then trihydroxy-isoflavanone comes into being due to isoflavanone synthase and finally, hydroxyisoflavanone dehydratase is responsible for the production of daidzein.

Based on our functional annotation findings, 34 contigs were predicted to represent seven enzymes essential to the biosynthesis of daidzein in *S.*

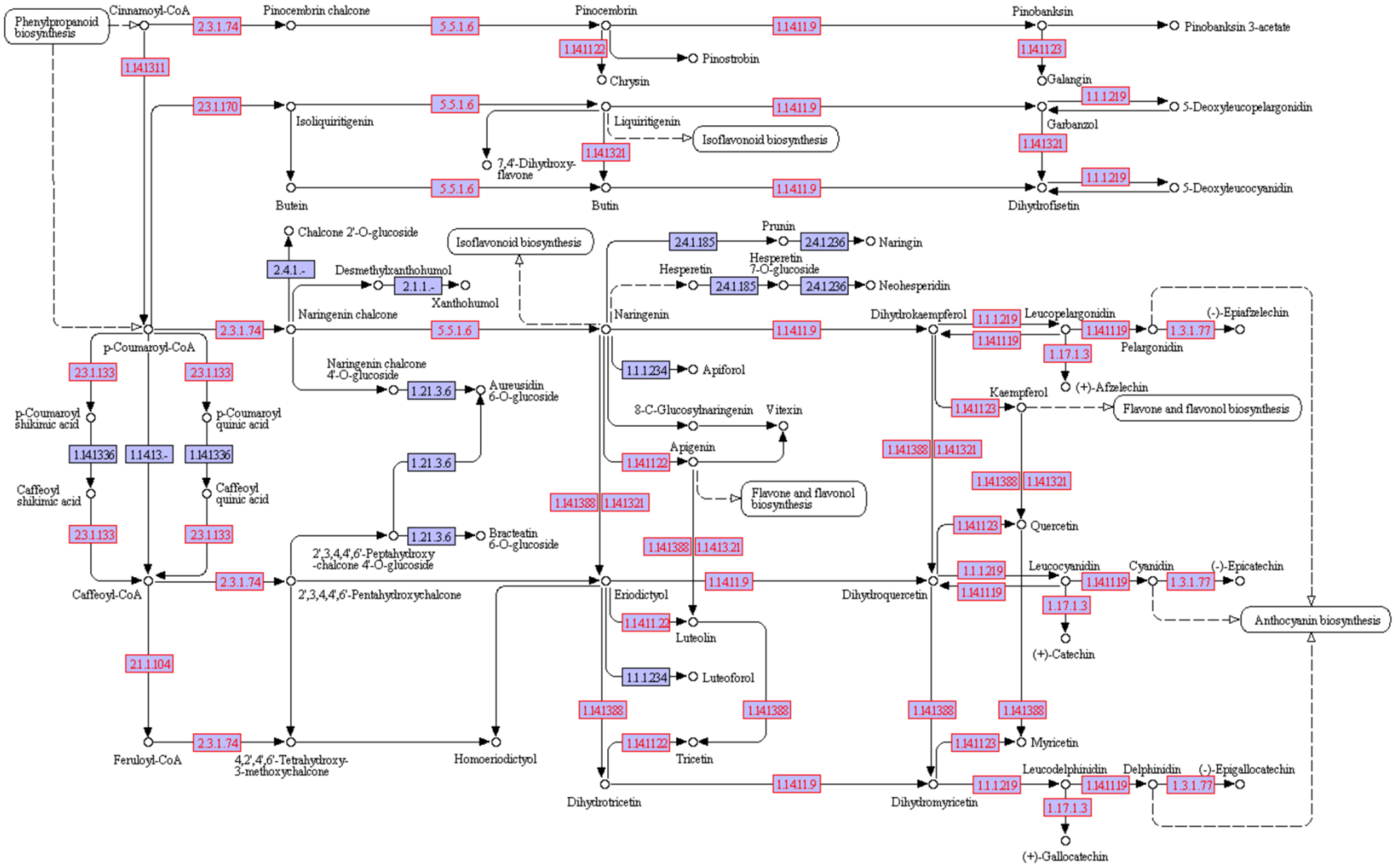

Figure 2.7 | Flavonoid biosynthetic pathway in *S. flavescens*. 18 EC numbers in red boxes were found.

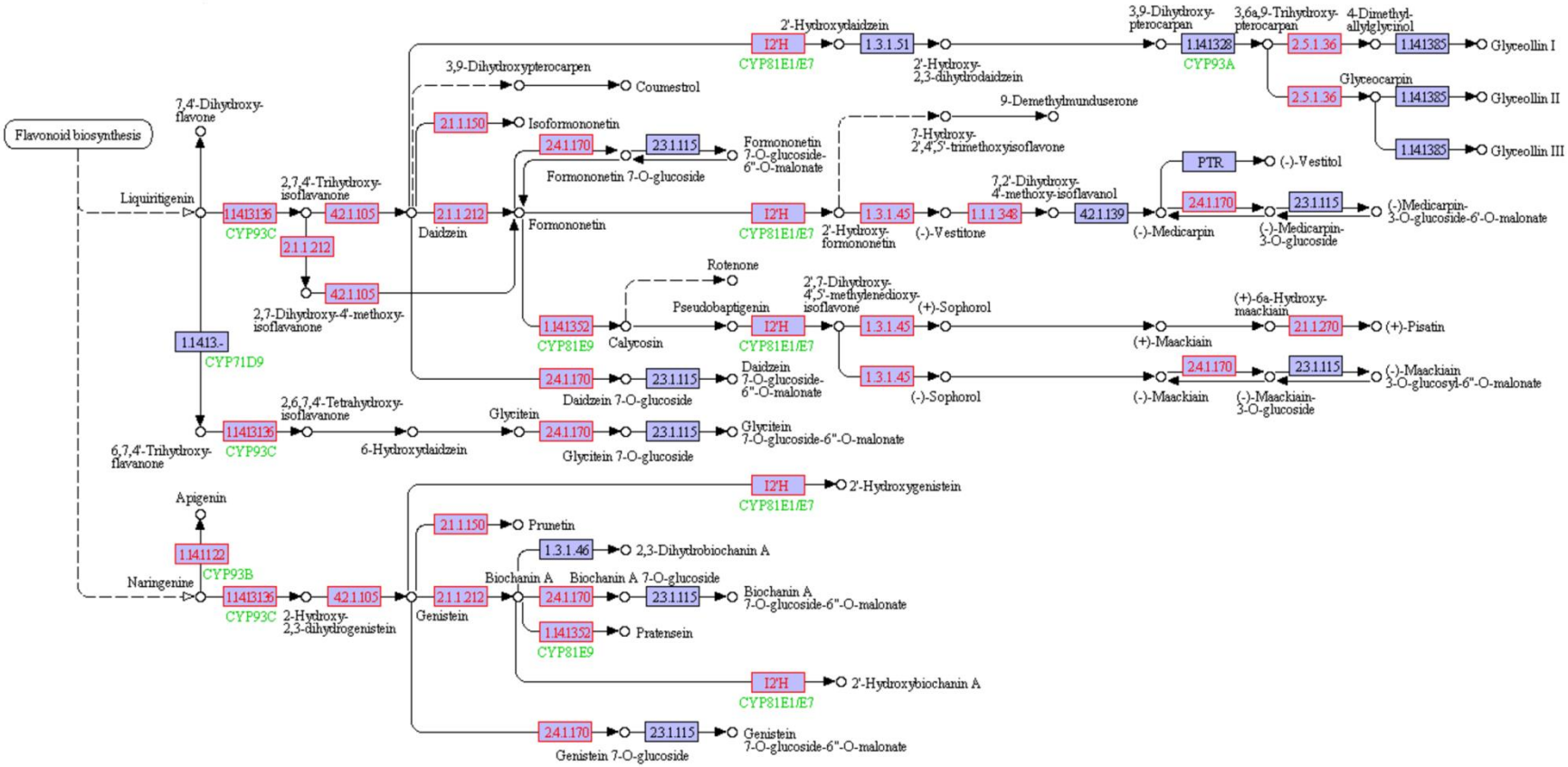

Figure 2.8 | *S. flavescens* isoflavonoid biosynthetic pathway.

The 13 enzyme commission numbers in red boxes were found in *S. flavescens* dataset.

flavescens. The number of contigs corresponding to each enzyme and the biosynthesis procedure are presented in **Figure 2.9**. In the annotation result, contig c_86767 had 97.54% similarity to the *L/ODC* which catalyzes lysine into cadaverine and thus initiates the first steps toward the final product quinolizidine alkaloids including matrine and oxymatrine (Bunsupa et al., 2012).

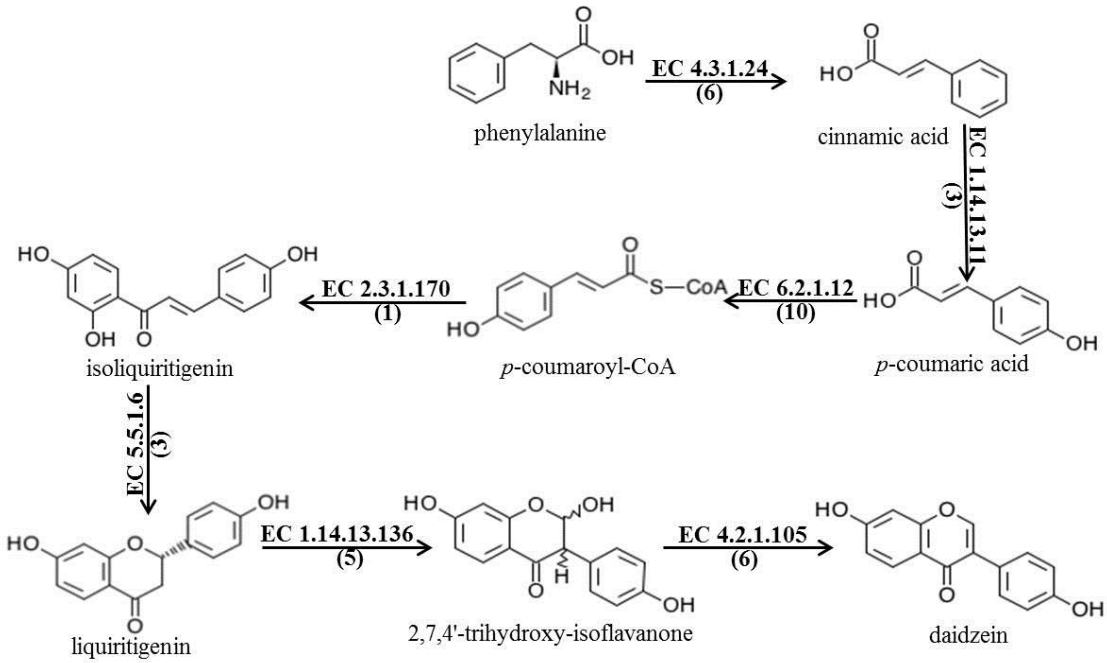

Figure 2.9 | Proposed daidzein biosynthetic pathway in *S. flavescens*. Each EC number is followed by the number of corresponding contigs in parentheses. EC 4.3.1.24: Phenylalanine ammonia-lyase, EC 1.14.13.11: *Trans*-cinnamate 4-monooxygenase, EC 6.2.1.12: 4-coumarate-CoA ligase, EC 2.3.1.170: Chalcone synthase, EC 5.5.1.6: Chalcone isomerase, EC 1.14.13.136: 2-hydroxyisoflavanone synthase, EC 4.2.1.105: 2-hydroxy- isoflavanone dehydratase.

By far, despite the sustained effort to crack the biosynthetic pathway regarding quinolizidine alkaloids, the identified genes underlying this process are very few. We focused on co-expression pattern related to c_86767 in order to hunt for candidate genes. According to RPKM values from all the studied organs, we calculate the Pearson correlation coefficients.

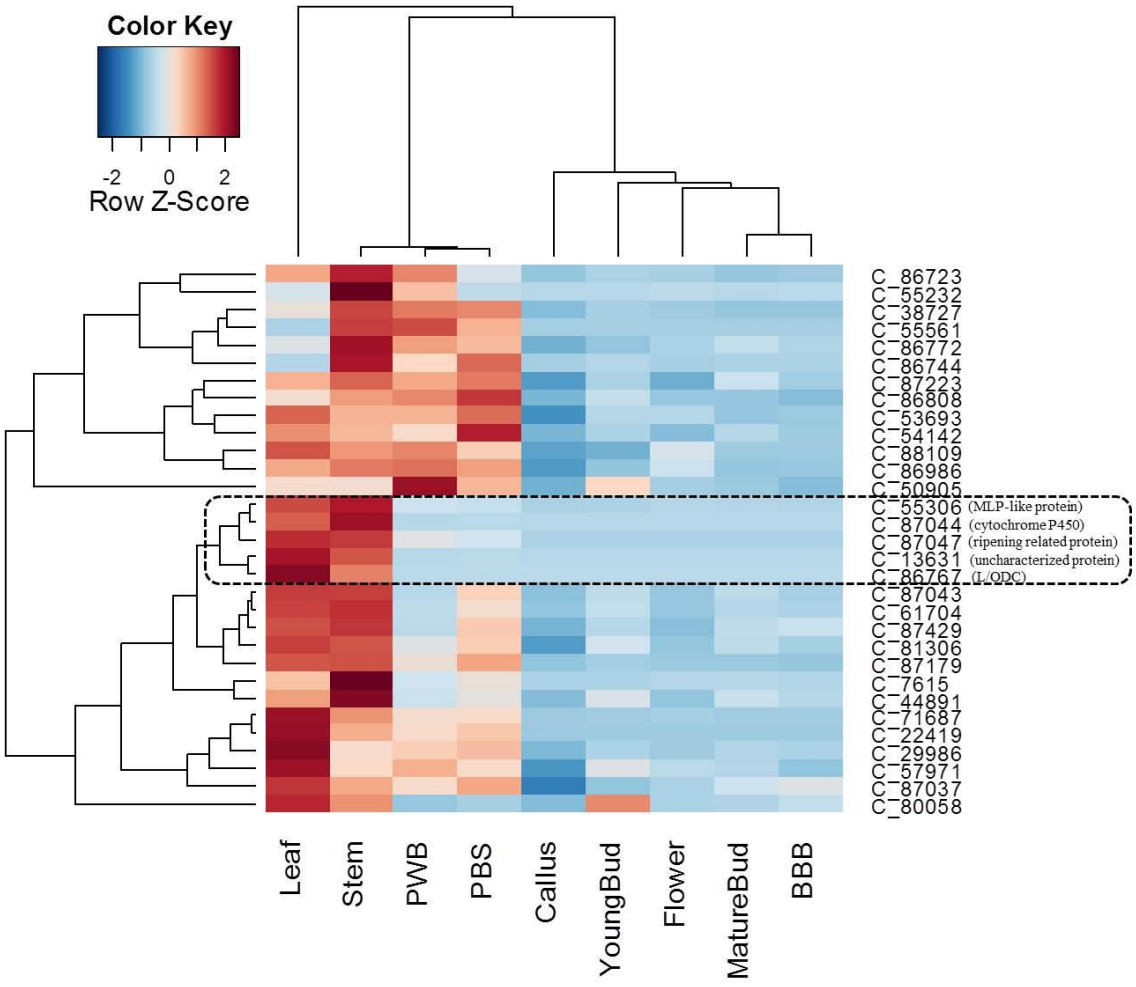

Figure 2.10 | Heatplot showing the expression profile for the 31 related contigs. The annotations for the transcripts highly related to the gene L/ODC are listed. C_55306 MLP-like protein, C_87044 cytochrome P450, C_87047 ripening related protein and C_13631 uncharacterized protein.

The heatplot (**Figure 2.10**) demonstrates several promising candidates including c_55306 (MLP-like protein), c_87044 (cytochrome P450) and c_87047 (ripening related protein). Such contigs shared very similar expression pattern across all the 9 tissues, suggesting their potential relationship in the process of quinolizidine alkaloid synthesis.

For contig c_86767 annotated as *L/ODC* gene, RPKM values of leaf and stem were as high as 553 and 321, respectively. Its expression pattern was further confirmed by semi-quantitative RT-PCR of *L/ODC* gene in the samples from leaf, stem and root of *S. flavescence*, using β-actin as the internal control gene (Hong et al., 2010), as shown in **Figure 2.11**.

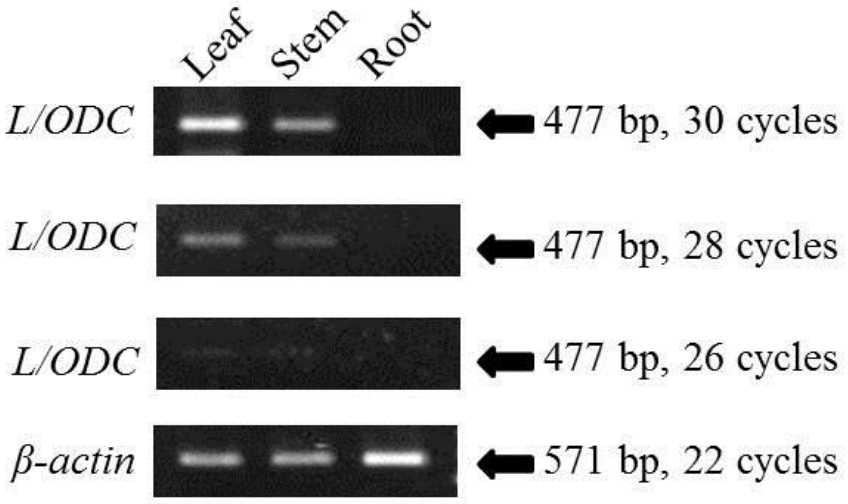

Figure 2.11 | Semi-quantitative RT-PCR validation for *L/ODC* gene.

Total RNA of kurara leaf, stem and root was extracted and then subjected to semi-quantitative RT-PCR using *S. flavescens* β-actin as an internal control.

PCR amplification was carried out with three different cycle numbers (30, 28 and 26 cycles) to verify the linearity of semi-quantitative analysis. The highest expression was seen in leaf followed by stem. No detectable expression was observed in root in this condition. These RT-PCR results verified the expression pattern deduced from RPKM values provided by RNA-seq analysis.

Furthermore, it was also consistent with the preferential *L/ODC* expression in the leaf of *Lupinus angustifolius* (Bunsupa et al., 2012). In all, accumulation of the final product such as matrine and oxymatrine would be high in other organs including root, but the initial steps for quinolizidine alkaloid biosynthesis feature in green parts of *S. flavescens*, such as leaf and stem as suggested previously (Saito and Murakoshi, 1995).

2.3.8 Cloning of one *S. flavescens* CYP86A24-like gene

Since callus of *S. flavascens* does not produce quinolizidine alkaloids (Saito et al., 1989), not to mention matrine or oxymatrine, while stem and leaf of kurara do accumulate considerable amount of matrine and oxymatrine (Wang et al., 2008), we searched the pool of DE contigs between callus and leaf / stem. 139 transcripts were found to be highly expressed in leaf and stem and 5 of them were annotated as monooxygenase or cytochrome P450. We focused on contig C_84372 and C_87778 and the BLAST result suggested these two contigs may belong to the different parts of a single gene. We performed the experiment to clone the full-length sequence and the similarity with CYP86A24 from *Glycine*

max is over 80%.

Although the functional identification for CYP86A24 could not be found, according to the previous reports (Hofer et al., 2008; Serra, et al., 2009), AtCYP86A1 and QsCYP86A32 are fatty acid omega-hydroxylases that are potentially involved in suberin biosynthesis.

The sequence for *S. flavescens* CYP86A24-like gene is presented in **Supplementary 2.2**. And the phylogenetic tree for the selected genes (**Figure 2.12**) suggests SfCYP450 may also have the function as a fatty acid omega-hydroxylase. Functional enzyme assays are now being undertaken.

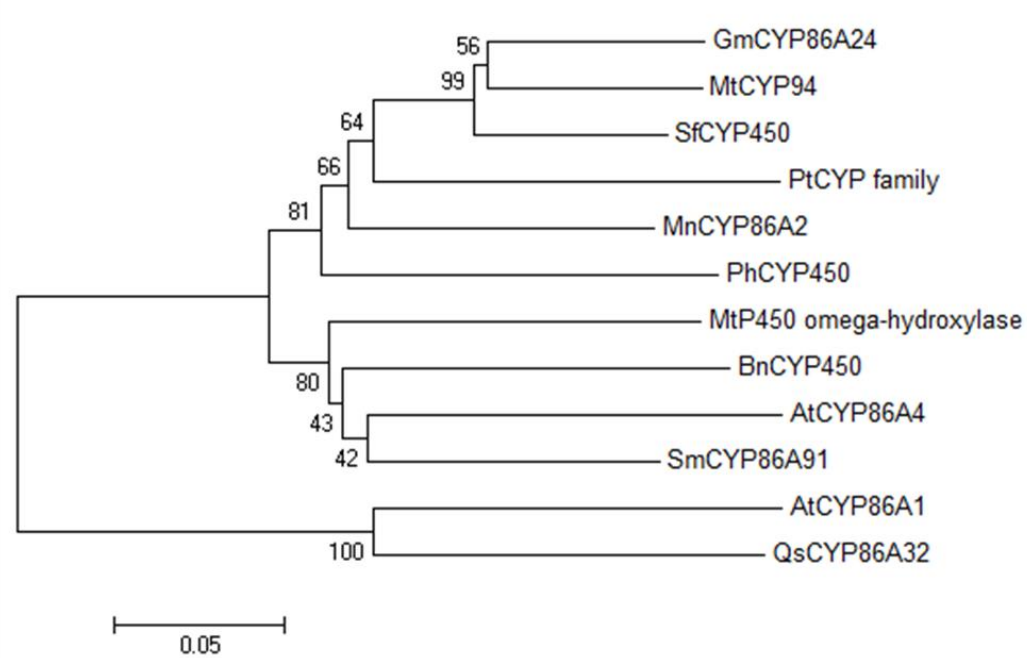

Figure 2.12 | Phylogenetic tree for selected genes related to SfCYP450 (SfCYP86A24-like). Gm *Glycine max*, Mt *Medicago truncatula*, Sf *Sophora flavescens*, Pt *populus trichocarpa*, Mn *Morus notabilis*, Ph *Petunia hybrid*, Bn *Brassica napus*, At *Arabidopsis thaliana*, Sm *Salvia miltiorrhiza* and Qs *Quercus suber*.

2.4 Discussion

S. flavescens has remarkable pharmaceutical attributes, including treating viral hepatitis, viral myocarditis, gastrointestinal hemorrhage and skin diseases (such as psoriasis and eczema).

Numerous studies have reported the isolation and pharmacological action of the bioactive components in *S. flavescens*. Despite the importance and wide application of this medicinal plant, very few studies focused on its genetic profile. To generate overall deep transcriptome data for this plant will be helpful in discovering its underlying mechanism for the production of quinolizidine alkaloids. In this study, we utilized more than 200 million fastq format reads resulted from 9 tissues of *S. flavescens* to perform RNA-Seq analysis. 83,325 representative contigs were obtained and the key enzymes involved in the biosynthetic pathways of active compounds were retrieved.

As to differentially expressed genes, NOISeq-sim output (**Figure 2.5**) showed DE genes identified by pair-wise comparison between callus and the rest 8 tissues were more than that of any other individual comparison. This is quite reasonable because in contrast with other organs of *S. flavescens*, callus consists of undifferentiated photoautotrophic cells.

And in the slightly different situations as mature bud and bud right before blossom, not a single DE gene could be detected. Callus of *S. flavescens* does not accumulate quinolizidine alkaloids, which allows for the responsible genes to pop up by comparing it with other capable organs. By measuring the changes of matrine and oxymatrine in different growth

stages and organs (Zhang et al., 2008 and Wang et al., 2008) of *S. flavescens* by HPLC, the concentration of above-mentioned compounds in root and seeds was 3- to 6-fold higher than that of leaf, stem and flower. Nevertheless, such reports also showed leaf, stem and flower did accumulate considerable quantity of matrine and oxymatrine. These observations suggest that the gene(s) responsible for the formation of such characteristic compounds should lie in the list of differentially expressed genes (**Figure 2.4**) and the alkaloids may translocate from the sites of *de novo* biosynthesis to the different sites of final accumulation.

Feeding studies illustrated the very first step for quinolizidine alkaloids biosynthesis. It concerns the decarboxylation of L-Lysine into cadaverine by catalysis of *L/ODC* (EC 4.1.1.18). With the presence of copper amine oxidase (*CuAO*, EC 1.4.3.22), oxidative deamination of cadaverine produces 5-aminopentanal that spontaneously cyclizes to Δ^1-piperideine Schiff base (Bunsupa et al., 2012; Ma and Gang, 2004).

The following steps to the final product matrine and oxymatrine still remain unknown. Based on co-expression pattern across the studied samples, we analyzed some candidate genes which were clustered into the same clade with the identified *L/ODC* gene. What's more, the genes suggested by the RNA-Seq data of *S. flavescens* are also the ones indicated by the deep transcriptome data of *Lupinus angustifolius*, a species of legume that accumulates considerable quinolizidine alkaloids (Bunsupa et al. unpublished).

Using enzyme commission numbers identified in *S. flavescens* dataset

to search KEGG for possible biosynthetic pathways retrieved 379 enzymes involved in biosynthesis of secondary metabolites. In addition to flavonoid biosynthetic pathway, valuable information about other pathways were also presented, such as monoterpenoid biosynthesis and steroid biosynthesis, which will contribute to the better understanding of the related mechanisms in plant kingdom.

In this study, the obtained putative novel genes which may underlie the pharmaceutical function of *S. flavescens* and the findings on several essential biosynthetic pathways may serve as a stepping-stone for further studies on this promising and time-honored medicinal plant.

General Discussion and Conclusions

Deep transcriptome data resulted from 5 tissues of *P. lobata* and 9 tissues of *S. flavescens* were achieved. Trimming process and assembly results showed the high quality of the raw fastq format reads and reliability of the assembled data.

In chapter one, GC content of *P. lobata* was found to be similar to that of Medicago truncatula and Glycine max. Gene ontology annotation and overrepresentation analysis using Fisher's exact test revealed certain GO terms may be related to its rapid and aggressive growth. For isoflavonoid biosynthetic pathway, 45 putative transcripts were retrieved and four contigs were chosen for biological validation using qRT-PCR. In general, the result was consistent with the RPKM ratio obtained by RNA-Seq analysis.

By HPLC, the accumulation of puerarin and daidzin in different tissues of *P. lobata* was measured. The underground parts contain the two compounds with varied concentration while leaf does not accumulate puerarin or daidzin. According to co-expression analysis, candidate genes for producing puerarin were proposed and because the production of puerarin may require transportation across different tissues, 41 expressed ABC transporters were also discussed.

In chapter 2, RNA-Seq data from 9 tissues were employed to search for putative genes involved in quinolizidine alkaloids biosynthesis. Due to the limited information concerning this pathway, genes co-expressed with *L/ODC* were studied and contigs annotated as MLP-like protein,

cytochrome P450 and ripening related protein may be involved in this pathway.

By focusing on the differentially expressed genes between callus and stem / leaf, we managed to pin down one putative cytochrome 450 gene with high similarity to CYP86A subfamily which shows fatty acid omega-hydroxylation capacity in the previous report (Hofer et al., 2008). The work to identify the function of this protein is now being undertaken.

We also used semi-quantitative PCR to measure the expression of *L/ODC* in leaf, stem and root of *S. flavescens*. The result suggested its highest expression in leaf, followed by stem, while there was no detectable expression in root.

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. (1991). Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science* 252, 1651-1656.

Anders, S. (2010). Htseq: analysing high-throughput sequencing data with python http://www-huber.embl.de/users/anders/HTSeq/.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25**,** 25-29. doi: 10.1038/75556.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57**,** 289-300.

Boiteau, L., Boivin, J., Liard, A., Quiclet-Sire, B., and Zard Z. (1998). A short synthesis of (±)-matrine. *Angew Chem Int Ed* 37, 1128-1131.

Bouque, V., Bourgaud, F., Nguyen, C., and Guckert, A. (1998). Production of daidzein by callus cultures of Psoralea species and comparison with plants. *Plant Cell Tissue and Organ Culture* 53**,** 35-40. doi: Doi 10.1023/A:1006057211490.

Brazier-Hicks, M., Evans, K.M., Gershater, M.C., Puschmann, H., Steel, P.G., and Edwards, R. (2009). The C-Glycosylation of Flavonoids in Cereals. *Journal of Biological Chemistry* 284**,** 17926-17934. doi: DOI 10.1074/jbc.M109.009258.

Bunsupa, S., Katayama, K., Ikeura, E., Oikawa, A., Toyooka, K., Saito, K., and Yamazaki, M. (2012). Lysine decarboxylase catalyzes the first step of quinolizidine alkaloid biosynthesis and coevolved with alkaloid production in leguminosae. *Plant Cell*, 24, 1202-1216.

Bunsupa, S., Yamazaki, M., and Saito, K. (2012). Quinolizidine alkaloid biosynthesis: recent advances and future prospects. *Front Plant Sci*, 3, 239.

Carai, M.a.M., Agabio, R., Bombardelli, E., Bourov, I., Gessa, G.L., Lobina, C., Morazzoni, P., Pani, M., Reali, R., Vacca, G., and Colombo, G. (2000). Potential use of medicinal plants in the treatment of alcoholism. *Fitoterapia* 71**,** S38-S42.

Chen, G., Li, R., Shi, L., Qi, J., Hu, P., Luo, J., Liu, M., Shi, T. (2011). Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. *BMC genomics*, 12, 590.

Chen, G., Zhang, J., and Jiannong, Y. (2001). Determination of puerarin, daidzein and rutin in *Pueraria lobata* (Wild.) Ohwi by capillary electrophoresis with electrochemical detection. *J Chromatogr A* 923**,** 255-262.

Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC*

Bioinformatics 12**,** 35. doi: 10.1186/1471-2105-12-35.

Chen, X., Yi, C., Yang, X., and Wang, X. (2004). Liquid chromatography of active principles in *Sophora flavescens* root. *J Chromatogr B Analyt Technol Biomed Life Sci*, 812, 149-163.

Cherdshewasart, W., Subtang, S., and Dahlan, W. (2007). Major isoflavonoid contents of the phytoestrogen rich-herb Pueraria mirifica in comparison with *Pueraria lobata*. *J Pharm Biomed Anal* 43**,** 428-434. doi: 10.1016/j.jpba.2006.07.013.

Cherdshewasart, W., and Sutjit, W. (2008). Correlation of antioxidant activity and major isoflavonoid contents of the phytoestrogen-rich *Pueraria mirifica* and *Pueraria lobata* tubers. *Phytomedicine* 15**,** 38-43. doi: 10.1016/j.phymed.2007.07.058.

Falcone Ferreyra, M.L., Rius, S.P., and Casati, P. (2012). Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Front Plant Sci* 3**,** 222. doi: 10.3389/fpls.2012.00222.

Ferreyra, M.L.F., Rodriguez, E., Casas, M.I., Labadie, G., Grotewold, E., and Casati, P. (2013). Identification of a Bifunctional Maize C- and O-Glucosyltransferase. *Journal of Biological Chemistry* 288**,** 31678-31688. doi: DOI 10.1074/jbc.M113.510040.

Franca, L.T.C., Carrilho, E., and Kist, T.B.L. (2002). A review of DNA sequencing techniques. *Q Rev Biophys* 35, 169-200.

Follak, S. (2011). Potential distribution and environmental threat of *Pueraria lobata*. *Central European Journal of Biology* 6**,** 457-469.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for

clustering the next-generation sequencing data. *Bioinformatics* 28**,** 3150-3152. doi: 10.1093/bioinformatics/bts565.

Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* 19**,** 521-532. doi: 10.1101/gr.074906.107.

Funaya, N., and Haginaka, J. (2012). Matrine- and oxymatrine-imprinted monodisperse polymers prepared by precipitation polymerization and their applications for the selective extraction of matrine-type alkaloids from *Sophora flavescens* Aiton. *J Chromatogr A*, 1248, 18-23.

Gaines, T.A., Lorentz, L., Figge, A., Herrmann, J., Maiwald, F., Ott, M.C., Han, H., Busi, R., Yu, Q., Powles, S.B., and Beffa, R. (2014). RNA-Seq transcriptome analysis to identify genes involved in metabolism-based diclofop resistance in Lolium rigidum. *Plant J* 78**,** 865-876. doi: 10.1111/tpj.12514.

Gao, T., Sun, Z., Yao, H., Song, J., Zhu, Y., Ma, X., and Chen, S. (2011). Identification of Fabaceae plants using the DNA barcode *matK*. *Planta Med*, 77, 92-94.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29**,** 644-652. doi: 10.1038/nbt.1883.

Graham, T.L. (1991). Flavonoid and Isoflavonoid Distribution in

Developing Soybean Seedling Tissues and in Seed and Root Exudates. *Plant Physiology* 95**,** 594-603. doi: Doi 10.1104/Pp.95.2.594.

He, X., Blount, J.W., Ge, S., Tang, Y., and Dixon, R.A. (2011). A genomic approach to isoflavone biosynthesis in kudzu (*Pueraria lobata*). *Planta* 233**,** 843-855. doi: 10.1007/s00425-010-1344-1.

He, X.Z., Li, W.S., Blount, J.W., and Dixon, R.A. (2008). Regioselective synthesis of plant (iso)flavone glycosides in Escherichia coli. *Appl Microbiol Biotechnol* 80**,** 253-260. doi: 10.1007/s00253-008-1554-7.

Hofer, R., Briesen, I., Beck, M., Pinot, F., Schreiber, L., and Franke, R., (2008). The *Arabidopsis* cytochrome P450 *CYP86A1* encodes a fatty acid ω-hydroxylase involved in suberin monomer biosynthesis. *J Exp Bot* 59, 2347-2360. doi:10.1093/jxb/ern101.

Hong, M.H., Lee, J.Y., Jung, H., Jin, D.H., Go, H.Y., Kim, J.H., Jang, B.H., Shin, Y.C., and Ko, S.G. (2009). *Sophora flavescens* Aiton inhibits the production of pro-inflammatory cytokines through inhibition of the NF kappaB/IkappaB signal pathway in human mast cell line (HMC-1). *Toxicol In Vitro,* 23, 251-258.

Hong, S.M., Bahn, S.C., Lyu, A., Jung, H.S., and Ahn, J.H. (2010). Identification and testing of superior reference genes for a starting pool of transcript normalization in Arabidopsis. *Plant Cell Physiol* 51**,** 1694-1706. doi: 10.1093/pcp/pcq128.

Jung, W., Yu, O., Lau, S.M., O'keefe, D.P., Odell, J., Fader, G., and Mcgonigle, B. (2000). Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nat*

Biotechnol 18**,** 208-212. doi: 10.1038/72671.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40**,** D109-114. doi: 10.1093/nar/gkr988.

Kawaguchi, R., and Bailey-Serres, J. (2005). mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Research* 33**,** 955-965. doi: Doi 10.1093/Nar/Gki240.

Kerscher, F., and Franz, G. (1987). Biosynthesis of Vitexin and Isovitexin - Enzymatic-Synthesis of the C-Glucosylflavones Vitexin and Isovitexin with an Enzyme Preparation from Fagopyrum-Esculentum M Seedlings. *Zeitschrift Fur Naturforschung C-a Journal of Biosciences* 42**,** 519-524.

Keung, W.M., and Vallee, B.L. (1998). Kudzu root: an ancient Chinese source of modern antidipsotropic agents. *Phytochemistry* 47**,** 499-506.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 6**,** 291-295. doi: 10.1038/nmeth.1311.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). Genome Project Data Processing S: The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078-2079.

Li, L.Q., Li, X.L., Wang, L., Du, W.J., Guo, R., Liang, H.H., Liu, X., Liang,

D.S., Lu, Y.J., Shan, H.L., and Jiang H.C. (2012). Matrine inhibits breast cancer growth via miR-21/PTEN/Akt pathway in MCF-7 cells. *Cell physiol biochem,* 30, 631-641.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25**,** 1966-1967. doi: 10.1093/bioinformatics/btp336.

Lindgren, C.J., Castro, K.L., Coiner, H.A., Nurse, R.E., and Darbyshire, S.J. (2013). The Biology of Invasive Alien Plants in Canada. 12. *Pueraria montana* var. *lobata* (Willd.) Sanjappa & Predeep. *Canadian Journal of Plant Science* 93**,** 71-95. doi: Doi 10.4141/Cjps2012-128.

Ling, J.Y., Zhang, G.Y., Cui, Z.J., Zhang, C.K. (2007). Supercritical fluid extraction of quinolizidine alkaloids from *Sophora flavescens* Ait. and purification by high-speed counter-current chromatography. *J Chromatogr A,* 1145, 123-127.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *J Biomed Biotechnol* 2012, 1-11. doi: 10.1155/2012/251364

Ma, X. and Gang, D.R. (2004). The Lycopodium alkaloids. *Nat Prod Rep*, 21, 752-772.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011**,** bar009. doi: 10.1093/database/bar009.

Maria, L., Falcone, F., Sebastian, P.R., and Paula, C. (2012). Flavonoids: Biosynthesis, Biological Functions, and Biotechnological Applications.

Front Plant Sci 2012, 222. doi:10.3389/fpls.2012.00222

Martin, J.A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet* 12, 671-682. doi:10.1038/nrg3068

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., Raney, B.J., Pohl, A., Malladi, V.S., Li, C.H., Lee, B.T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R.A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B.M., Fujita, P.A., Dreszer, T.R., Diekhans, M., Cline, M.S., Clawson, H., Barber, G.P., Haussler, D., and Kent, W.J. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res,* 41, D64-D69.

Miadokova, E. (2009). Isoflavonoids - an overview of their biological activities and potential health benefits. *Interdiscip Toxicol* 2**,** 211-218. doi: 10.2478/v10102-009-0021-3.

Michael, T.P. and Jackson, S. (2013). The first 50 plant genomes. *Plant Genome-Us,* 6.

Miyazawa, M., Sakano, K., Nakamura, S., and Kosaka, H. (2001). Antimutagenic activity of isoflavone from *Pueraria lobata*. *J Agric Food Chem* 49**,** 336-341.

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

Muranaka, T., and Saito, K. (2013). Phytochemical Genomics on the Way. *Plant and Cell Physiology* 54**,** 645-646. doi: Doi 10.1093/Pcp/Pct058.

Nagai, N. (1889). Study on *Sophora flavescens*. *Yakugaku Zasshi*, 84, 54-87.

Okuda, S., Yoshimoto, M., Tsuda, K., and Utzugi, N. (1966). On the absolute configuration of matrin. *Chem Pharm Bull (Tokyo)*, 14, 314-318.

Ralston, L., Subramanian, S., Matsuno, M., and Yu, O. (2005). Partial Reconstruction of Flavonoid and Isoflavonoid Biosynthesis in Yeast Using Soybean Type I and Type II Chalcone Isomerases. *Plant Physiol* 137, 1375-1388.

Ramilowski, J.A., Sawai, S., Seki, H., Mochida, K., Yoshida, T., Sakurai, T., Muranaka, T., Saito, K., and Daub, C.O. (2013). Glycyrrhiza uralensis transcriptome landscape and study of phytochemicals. *Plant Cell Physiol* 54**,** 697-710. doi: 10.1093/pcp/pct057.

Saito, K. (2013). Phytochemical genomics - a new trend. *Current Opinion in Plant Biology* 16**,** 373-380. doi: DOI 10.1016/j.pbi.2013.04.001.

Saito, K. and Murakoshi, I. (1995). Chemistry, biochemistry and chemotaxonomy of lupine alkaloids in the leguminosae. *Studies in Natural Products Chemistry*. (Atta ur R Vol. Volume 15, Part C. Elsevier, 519-549.

Saito, K., Yamazaki, M., Yamakawa, K., Fujisawa, S., Takamatsu, S., Kawaguchi, A., and Murakoshi, I. (1989). Lupin alkaloids in tissue-culture of *Sophora flavescens var angustifolia* - Greening induced production of matrine. *Chem Pharm Bull (Tokyo)*, 37, 3001-3004.

Saito, K., Yonekura-Sakakibara, K., Nakabayashi, R., Higashi, Y., Yamazaki, M., Tohge, T., and Fernie, A.R. (2013). The flavonoid biosynthetic pathway in Arabidopsis: structural and genetic diversity. *Plant Physiol Biochem* 72**,** 21-34. doi: 10.1016/j.plaphy.2013.02.001.

Sasaki, K., Tsurumaru, Y., Yamamoto, H., and Yazaki, K. (2011). Molecular characterization of a membrane-bound prenyltransferase specific for isoflavone from *Sophora flavescens*. *J Biol Chem*, 286, 24125-24134.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* 3**,** 1101-1108.

Serra, O., Soler, M., Hohn, C., Sauveplane, V., Pinot, F., Franke, R., Schreiber, L., Prat, S., Molinas, M., and Figueras, M. (2009). CYP86A33-Targeted Gene Silencing in Potato Tuber Alters Suberin Composition, Distorts Suberin Lamellae, and Impairs the Periderm's Water Barrier Function. *Plant Physiol* 2009, 1050-1060.

Serres-Giardi, L., Belkhir, K., David, J., and Glemin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *Plant Cell* 24**,** 1379-1397. doi: DOI 10.1105/tpc.111.093674.

Shen, Y., Feng, Z.M., Jiang, J.S., Yang, Y.N., Zhang, P.C. (2013). Dibenzoyl and isoflavonoid glycosides from *Sophora flavescens*: inhibition of the cytotoxic effect of D-galactosamine on human hepatocyte HL-7702. *J Nat Prod*, 76, 2337-2345.

Shibata, S. and Sankawa, U. (1963). Biosynthesis of matrine. *Chem Ind*, 1161-1162.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol,

I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* 19**,** 1117-1123. doi: 10.1101/gr.089532.108.

Sun, M., Cao, H., Sun, L., Dong, S., Bian, Y., Han, J., Zhang, L., Ren, S., Hu, Y., Liu, C., Xu, L., and Liu, P. (2012). Antitumor activities of kushen: literature review. *J Evid Based Complementary Altern Med: eCAM*, 373219.

Steele, C.L., Gijzen, M., Qutob, D., and Dixon, R.A. (1999). Molecular characterization of the enzyme catalyzing the aryl migration reaction of isoflavonoid biosynthesis in soybean. *Archives of biochemistry and biophysics* 367**,** 146-150.

Takahashi, H., Morioka, R., Ito, R., Oshima, T., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Dynamics of time-lagged gene-to-metabolite networks of Escherichia coli elucidated by integrative omics approach. *OMICS* 15**,** 15-23. doi: 10.1089/omi.2010.0074.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res* 21**,** 2213-2223. doi: 10.1101/gr.124321.111.

Teng, Y., Cui, H., Yang, M., Song, H., Zhang, Q., Su, Y., and Zheng, J. (2009). Protective effect of puerarin on diabetic retinopathy in rats. *Mol Biol Rep* 36**,** 1129-1133. doi: 10.1007/s11033-008-9288-2.

Tian, A.G., Wang, J., Cui, P., Han, Y.J., Xu, H., Cong, L.J., Huang, X.G., Wang, X.L., Jiao, Y.Z., Wang, B.J., Wang, Y.J., Zhang, J.S., and Chen, S.Y. (2004). Characterization of soybean genomic features by analysis of its expressed sequence tags. *Theor Appl Genet* 108**,** 903-913. doi:

10.1007/s00122-003-1499-2.

Wang, L.L., Ma, P.Q., and Pei, C.J. (2008). Determination of oxymatrine content in different organs of *Sophora flavescens* Ait. by HPLC. *Journal of Anhui Agri Sci*, 36, 5691-5692.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10**,** 57-63. doi: 10.1038/nrg2484.

Wong, K.H., Li, G.Q., Li, K.M., Razmovski-Naumovski, V., and Chan, K. (2011). Kudzu root: traditional uses and potential medicinal benefits in diabetes and cardiovascular diseases. *J Ethnopharmacol* 134**,** 584-607. doi: 10.1016/j.jep.2011.02.001.

Yamamoto, H., Kawai, S., Mayumi, J., Tanaka, T., Iinuma, M., and Mizuno, M. (1991). Prenylated flavanone production in callus-cultures of *Sophora flavescens var angustifolia*. *Z Naturforsch. C*, 46, 172-176.

Ye, J., Fang, L., Zheng, H.K., Zhang, Y., Chen, J., Zhang, Z.J., Wang, J., Li, S.T., Li, R.Q., Bolund, L., and Wang, J. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* 34**,** W293-W297. doi: Doi 10.1093/Nar/Gkl1031.

Zerbino, D.R. and Birney, E. (2008). Velvet: Algorithms for *de novo* Short Read Assembly Using de Bruijn Graphs. *Genome Res* 2008, 821-829. doi:10.1101/gr.074492.107

Zhang, S.R., Ji, Y., and Lin, H.M. (2008). Changes of oxmatrine and matrine in the development of *Sophora flavescens* Ait. *Pratacultural Science*, 25, 41-45.

Acknowledgements

I would like to express the deepest respect and gratitude to Professor Kazuki Saito, my supervisor, whose broad and profound knowledge in almost every scope related to plant molecular biology has been inspiring me since I was kindly accepted as a student in Department of Molecular Biology and Biotechnology, Chiba University. His understanding and encouraging instructions guided me throughout the PhD course.

I wish to convey my sincere thanks to Associate Professor Hiroki Takahashi (Medical Mycology Research Center, Chiba University) for his kind and patient guidance in experimental planning and scattered details concerning analysis on bioinformatics.

I wish to express my warm appreciation to Associate Professor Mami Yamazaki and Assistant Professor Naoko Yoshimoto for their detailed and constructive comments and the essential support for this work.

I also wish to thank Dr. Hideyuki Suzuki (Kazusa DNA Research Institute) and Professor Hirobumi Yamamoto (Toyo University) for their careful and considerate help in conducting the experimental procedure.

Many thanks are given to my dear friends in Department of Molecular Biology and Biotechnology who workd with me and provided selfishless suggestions on my project.

And above all, these would not have been possible without the scholarship generously awarded by Ministry of Education, Culture, Sports, Science and Technology, Japan.

I thank you all from the bottom of my heart!

List of Publications

Part of this thesis has been published in the following articles:

Han, R., Takahashi, H., Nakamura, M., Yoshimoto, N., Suzuki, H., Shibata, D., Yamazaki, M., and Saito K. (2015). Transcriptomic landscape of *Pueraria lobata* demonstrates potential for phytochemical study. *Front Plant Sci* 6:426. doi: 10.3389/fpls.2015.00426.

Han, R., Takahashi, H., Nakamura, M., Bunsupa, S., Yoshimoto, N., Yamamoto, H., Suzuki, H., Shibata, D., Yamazaki, M., and Saito K. (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.

Thesis Committee

This thesis entitled transcriptomic study on two medicinal plants *Pueraria lobata* and *Sophora flavescens* was submitted to the Graduate School of Pharmaceutical Sciences, Chiba University, Japan, in fulfillment of the requirements for the degree of Doctor of Philosophy (Ph. D.) and had been examined by the following thesis committee authorized by the Graduate School of Pharmaceutical Sciences of Chiba University.

Chairman: Naoto Yamaguchi , Ph. D., Professor of the Graduate School of Pharmaceutical Sciences, Chiba University.

Members: Motoyuki Itoh, Ph. D., Professor of the Graduate School of Pharmaceutical Sciences, Chiba University.

Members: Masami Ishibashi, Ph. D., Professor of the Graduate School of Pharmaceutical Sciences, Chiba University.

Supplementary 1.1 Over-represented GO terms resulted for *P. lobata*.

Probability	GO term	Gene_ontology_name	n_{11}	n_{12}	n_{21}	n_{22}
0.00E+00	GO:0005737	cytoplasm	1311	1152	13053	65992
0.00E+00	GO:0009507	chloroplast	1561	1376	12803	65768
0.00E+00	GO:0005829	cytosol	1225	462	13139	66682
0.00E+00	GO:0016021	integral component of membrane	1700	2082	12664	65062
0.00E+00	GO:0005634	nucleus	3346	3560	11018	63584
4.61E-239	GO:0005794	Golgi apparatus	649	356	13715	66788
8.24E-233	GO:0009570	chloroplast stroma	514	179	13850	66965
7.99E-177	GO:0009651	response to salt stress	429	186	13935	66958
3.44E-165	GO:0005774	vacuolar membrane	385	152	13979	66992
1.47E-160	GO:0009535	chloroplast thylakoid membrane	306	67	14058	67077
2.64E-103	GO:0009409	response to cold	251	108	14113	67036
3.69E-94	GO:0009611	response to wounding	224	92	14140	67052
4.11E-93	GO:0005802	trans-Golgi network	223	93	14141	67051
3.90E-92	GO:0006355	regulation of transcription, DNA-templated	469	572	13895	66572
1.84E-89	GO:0010200	response to chitin	250	142	14114	67002
1.85E-88	GO:0007030	Golgi organization	171	39	14193	67105
7.18E-88	GO:0016787	hydrolase activity	374	381	13990	66763
2.65E-87	GO:0003824	catalytic activity	359	352	14005	66792
8.27E-86	GO:0005768	endosome	228	118	14136	67026
2.69E-84	GO:0016192	vesicle-mediated transport	188	66	14176	67078
3.08E-84	GO:0005618	cell wall	278	206	14086	66938
2.06E-83	GO:0048046	apoplast	244	150	14120	66994
2.48E-79	GO:0003723	RNA binding	469	650	13895	66494
4.64E-79	GO:0019344	cysteine biosynthetic process	178	64	14186	67080
2.17E-74	GO:0042742	defense response to bacterium	216	131	14148	67013
1.21E-73	GO:0051788	response to misfolded protein	140	30	14224	67114
5.34E-73	GO:0009737	response to abscisic acid	230	159	14134	66985
1.03E-72	GO:0010027	thylakoid membrane organization	170	67	14194	67077
3.46E-72	GO:0048193	Golgi vesicle transport	159	54	14205	67090
5.54E-72	GO:0006094	gluconeogenesis	145	38	14219	67106
5.31E-71	GO:0006612	protein targeting to membrane	225	157	14139	66987
7.15E-71	GO:0015031	protein transport	186	94	14178	67050

9.86E-70	GO:0009853	photorespiration	127 23 14237 67121
3.32E-69	GO:0009414	response to water deprivation	196 114 14168 67030
2.28E-68	GO:0080129	proteasome core complex assembly	111 11 14253 67133
8.76E-66	GO:0010207	photosystem II assembly	130 32 14234 67112
8.08E-65	GO:0006886	intracellular protein transport	164 77 14200 67067
8.77E-65	GO:0006457	protein folding	204 141 14160 67003
7.03E-64	GO:0006508	proteolysis	362 484 14002 66660
5.43E-63	GO:0016567	protein ubiquitination	263 262 14101 66882
1.79E-61	GO:0019252	starch biosynthetic process	152 68 14212 67076
3.56E-61	GO:0005525	GTP binding	206 157 14158 66987
4.42E-61	GO:0009733	response to auxin	183 117 14181 67027
7.84E-59	GO:0005777	peroxisome	151 73 14213 67071
4.61E-58	GO:0050832	defense response to fungus	190 139 14174 67005
7.22E-58	GO:0009744	response to sucrose	138 57 14226 67087
2.71E-57	GO:0005507	copper ion binding	173 112 14191 67032
3.49E-57	GO:0009867	jasmonic acid mediated signaling pathway	171 109 14193 67035
9.04E-57	GO:0000139	Golgi membrane	156 86 14208 67058
1.18E-56	GO:0006635	fatty acid beta-oxidation	141 64 14223 67080
1.39E-56	GO:0009738	abscisic acid-activated signaling pathway	174 116 14190 67028
5.56E-55	GO:0009902	chloroplast relocation	100 18 14264 67126
9.71E-55	GO:0015995	chlorophyll biosynthetic process	113 32 14251 67112
1.27E-50	GO:0048767	root hair elongation	148 91 14216 67053
1.40E-41	GO:0009723	response to ethylene	124 79 14240 67065
1.76E-41	GO:0009644	response to high light intensity	125 81 14239 67063
1.50E-39	GO:0009408	response to heat	137 109 14227 67035
3.67E-39	GO:0003743	translation initiation factor activity	111 65 14253 67079
1.20E-36	GO:0006950	response to stress	182 217 14182 66927
1.34E-34	GO:0051301	cell division	154 165 14210 66979
2.79E-34	GO:0045454	cell redox homeostasis	99 60 14265 67084
3.34E-34	GO:0016049	cell growth	97 57 14267 67087
2.00E-30	GO:0009637	response to blue light	73 31 14291 67113

Supplementary 2.1 Over-represented GO terms for *S. flavescens*.

GO_term	Gene_ontology_name	*P*_value
GO:0005886	plasma membrane	0.00E+00
GO:0005737	cytoplasm	0.00E+00
GO:0009507	chloroplast	0.00E+00
GO:0005829	cytosol	0.00E+00
GO:0005634	nucleus	0.00E+00
GO:0016021	integral component of membrane	2.07E-258
GO:0009570	chloroplast stroma	8.37E-227
GO:0009506	plasmodesma	8.02E-189
GO:0046872	metal ion binding	8.69E-176
GO:0046686	response to cadmium ion	3.68E-167
GO:0005576	extracellular region	8.47E-167
GO:0009651	response to salt stress	7.50E-166
GO:0005794	Golgi apparatus	2.65E-163
GO:0009941	chloroplast envelope	5.30E-162
GO:0005739	mitochondrion	6.54E-156
GO:0016020	membrane	6.65E-144
GO:0005524	ATP binding	7.82E-139
GO:0005774	vacuolar membrane	2.11E-137
GO:0009535	chloroplast thylakoid membrane	1.88E-122
GO:0005618	cell wall	1.17E-117
GO:0005783	endoplasmic reticulum	8.31E-113
GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway	6.74E-106
GO:0005730	nucleolus	3.82E-99
GO:0048046	apoplast	2.29E-90
GO:0006098	pentose-phosphate shunt	4.88E-89
GO:0006096	glycolytic process	9.54E-86
GO:0006364	rRNA processing	1.52E-80
GO:0005840	ribosome	2.75E-77
GO:0051788	response to misfolded protein	7.98E-75
GO:0009853	photorespiration	1.15E-74
GO:0007030	Golgi organization	6.12E-74
GO:0003824	catalytic activity	2.11E-73

GO:0006351	transcription, DNA-templated	8.54E-73
GO:0006412	translation	3.37E-72
GO:0019344	cysteine biosynthetic process	5.01E-72
GO:0003735	structural constituent of ribosome	3.72E-70
GO:0005773	vacuole	9.96E-70
GO:0080129	proteasome core complex assembly	2.43E-69
GO:0003677	DNA binding	2.36E-68
GO:0010027	thylakoid membrane organization	1.67E-66
GO:0009409	response to cold	9.66E-64
GO:0006355	regulation of transcription, DNA-templated	1.34E-63
GO:0042742	defense response to bacterium	1.65E-63
GO:0009505	plant-type cell wall	2.73E-62
GO:0009793	embryo development ending in seed dormancy	6.03E-62
GO:0006511	ubiquitin-dependent protein catabolic process	3.70E-61
GO:0003700	sequence-specific DNA binding transcription factor activity	4.52E-61
GO:0006094	gluconeogenesis	4.88E-60
GO:0005507	copper ion binding	1.33E-59
GO:0015995	chlorophyll biosynthetic process	6.74E-59
GO:0045893	positive regulation of transcription, DNA-templated	2.51E-58
GO:0019252	starch biosynthetic process	2.80E-57
GO:0016126	sterol biosynthetic process	7.55E-57
GO:0015031	protein transport	7.62E-57
GO:0000166	nucleotide binding	6.49E-56
GO:0016117	carotenoid biosynthetic process	1.35E-54
GO:0006457	protein folding	2.87E-53
GO:0016787	hydrolase activity	1.11E-51
GO:0009737	response to abscisic acid	3.31E-51
GO:0010207	photosystem II assembly	4.85E-50
GO:0009744	response to sucrose	6.91E-47
GO:0048767	root hair elongation	1.38E-46
GO:0005802	trans-Golgi network	1.68E-46
GO:0001510	RNA methylation	4.03E-46
GO:0016192	vesicle-mediated transport	3.90E-45
GO:0005768	endosome	4.96E-45
GO:0009664	plant-type cell wall organization	1.11E-44

GO:0009414	response to water deprivation	1.89E-44
GO:0006833	water transport	2.33E-44
GO:0006612	protein targeting to membrane	5.58E-43
GO:0009902	chloroplast relocation	1.27E-42
GO:0010363	regulation of plant-type hypersensitive response	1.74E-42
GO:0006979	response to oxidative stress	1.96E-42
GO:0005525	GTP binding	2.58E-42
GO:0010200	response to chitin	4.32E-42
GO:0009611	response to wounding	5.03E-42
GO:0015979	photosynthesis	6.93E-42
GO:0048193	Golgi vesicle transport	7.29E-42
GO:0009408	response to heat	9.02E-42
GO:0005506	iron ion binding	1.20E-41
GO:0005777	peroxisome	2.13E-41
GO:0009640	photomorphogenesis	3.08E-41
GO:0009909	regulation of flower development	3.37E-41
GO:0006886	intracellular protein transport	1.49E-40
GO:0009750	response to fructose	1.70E-40
GO:0000023	maltose metabolic process	2.07E-40
GO:0006950	response to stress	3.74E-40
GO:0009965	leaf morphogenesis	4.25E-40
GO:0019761	glucosinolate biosynthetic process	4.86E-40
GO:0010075	regulation of meristem growth	5.15E-40
GO:0030244	cellulose biosynthetic process	7.38E-40
GO:0000139	Golgi membrane	7.94E-40
GO:0005789	endoplasmic reticulum membrane	1.64E-39
GO:0016491	oxidoreductase activity	3.09E-39
GO:0016567	protein ubiquitination	8.44E-39
GO:0034976	response to endoplasmic reticulum stress	1.20E-38
GO:0035304	regulation of protein dephosphorylation	1.62E-38
GO:0006635	fatty acid beta-oxidation	3.61E-38
GO:0030154	cell differentiation	5.92E-38
GO:0006972	hyperosmotic response	1.26E-37
GO:0020037	heme binding	1.68E-37
GO:0016049	cell growth	2.25E-37
GO:0009220	pyrimidine ribonucleotide biosynthetic process	2.28E-37

GO:0019013	viral nucleocapsid	2.68E-37
GO:0006816	calcium ion transport	2.68E-37
GO:0000502	proteasome complex	6.40E-37
GO:0006007	glucose catabolic process	8.47E-37
GO:0051301	cell division	1.15E-36
GO:0009534	chloroplast thylakoid	4.58E-36
GO:0009658	chloroplast organization	5.98E-36
GO:0043085	positive regulation of catalytic activity	1.65E-35
GO:0006508	proteolysis	4.59E-35
GO:0032440	2-alkenal reductase [NAD(P)] activity	6.41E-35
GO:0009733	response to auxin	9.18E-35
GO:0005975	carbohydrate metabolic process	3.10E-34
GO:0031348	negative regulation of defense response	3.49E-34
GO:0048481	ovule development	6.37E-34
GO:0006633	fatty acid biosynthetic process	7.00E-34
GO:0050832	defense response to fungus	9.16E-34
GO:0009845	seed germination	1.58E-33
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	2.04E-33
GO:0008152	metabolic process	2.60E-33
GO:0010228	vegetative to reproductive phase transition of meristem	4.74E-33
GO:0006623	protein targeting to vacuole	1.51E-32
GO:0009644	response to high light intensity	2.18E-32
GO:0030529	ribonucleoprotein complex	2.23E-32
GO:0016132	brassinosteroid biosynthetic process	4.54E-32
GO:0005759	mitochondrial matrix	5.33E-32
GO:0006626	protein targeting to mitochondrion	3.01E-31
GO:0006499	N-terminal protein myristoylation	4.85E-31
GO:0006084	acetyl-CoA metabolic process	5.06E-31
GO:0006606	protein import into nucleus	8.01E-31

Supplementary 2.2 Fasta format sequence of CYP86A24-like gene

>SfCYP86A24-like

ATGGATGCATCAACGGCTTTTATGATCCTATCAGCCATGGGAGGCTATTTAATATGGTTCTCCTT
CATCTCGCGGTCACTGAGAGGTCCACGTGTCTGGCCCCTATTGGGTAGTCTCCCAGGTCTCAT
CCAACACGCCAACCGCATGCACGACTGGATCTCAGACAACCTCCGCGCGTGTGGCGGCACGT
ACCAAACCTGCATCTGTCCCATTCCCTTCCTCGCCAGAAAACAGGGTCTCGTGACCGTCACGT
GCGACCCCAAGAACCTCGAGCACATCCTCAAGCTCCGCTTCGACAACTACCCCAAGGGTCCG
ACGTGGCAGGCAGTATTCCACGACTTGCTCGGAGATGGCATTTTCAATTCAGATGGTGACACG
TGGCTGTTCCAGCGCAAGACCGCCGCGCTGGAATTCACCACCGCACCCTGCGCCAAGCCATG
GCCCGCTGGGTGAGCCGAGCCATCAAGCACAGGTTCTGTCCCATCTTAGCCGCCGCACAGCA
TGATCAGAAGTCTGTCGACCTCCAGGACCTGCTGCTTCGGCTCACTTTCGATAACATATGCGG
CTTGGCTTTCGGGCAAGACCCACAAACACTTGACGTGGGCCTACCCGAAAACAAGTTCGCAT
TGTCTTTCGACCGTGCAACCGAAGCCACGCTGCAACGCTTCATCTTGCCCGAAATTGTTTGGA
AGTTTAAGAAATGGCTTGGACTCGGGATGGAAGTGAGCCTGACCCAAAGCCTCAGACACATT
GATAAGTACCTTTCCAACATCATCAACACGCGCAAGCTTGAGCTGGTGGAAAAACAACAAGT
CATTGGTGCTGGTGGGGCCACCCATGATGACCTGTTATCTCGGTTCATGAAAAAGAAGGAATC
CTACTCAAACGAGTTCCTCCAACACGTGGCACTCAACTTCATCCTAGCTGGACGTGACACATC
ATCGGTGGCACTCAGCTGGTTCTTCTGGCTATGCATCCTAAATCCCAGCGTAGAGGAAAAGAT
CTTGATCGAGCTCTGCACCGTTCTGATGGAGACACGTGGCGGTGACGTGTCAAAGTGGGTCG
ACGAGCCTCTAGTGTTCGAGGAGGTTGACCGGTTGGTGTACCTGAAGGCCGCACTGTCGGAG
ACGCTGCGGCTTTACCCGTCGGTGCCGGAGGATTCGAAGCACGTGGTGAACGACGACGTTTT
GCCGAACGGGACGTTCGTTCCGGCGGGTTCAGCGGTTACCTATTCCATTTACAGCATCGGGAG
GATGAAGTTCATTTGGGGAGAGGACTGCCTGGAGTTCAAGCCGGAGCGGTGGCTCTCCGCCG
ACGGGAAACAGATTCAGGTGCATGATTCTTACAAATTCGTTTCGTTCAATGCGGGGCCCAGGA
TTTGCTTGGGGAAGGACTTGGCTTACTTGCAGATGAAGTCCATAGCGGCGGCGGTGCTGCTCC
GCCACCGCCTCACGGTGGCGCCGGGACACCGCGTGGAGCAGAAGATGTCGCTGACGCTGTTC
ATGAAGTATGGGCTAAAGGTGAACGTGCACCCTAGGGATCTAAGGCCGGTGTTGGAAAAGAT
AAAAAGCAAGGTTGAGTCGTGTGGTAAAGAAGCTCTCAGTAATAATGGTAATATGGACGGGG
TTGAAATGGTTGCTGCTGATGCT