INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
19–1344/2020/23–2–302–310
DOI: 10.17957/IJAB/15.1289
http://www.fspublishers.org

Full Length Article

De novo Sequencing and Transcriptome Analysis of *Prunella vulgaris* during Development: A Cross-Databases Comparison

Zanzan Li[1†], Yuhang Chen[2†], Qiaosheng Guo[1*], Changlin Wang[1], Liping Cao[2], Qin Qin[2], Miao Zhao[2] and Chen Li[2]
[1]*Institute of Chinese Medicinal Materials, Nanjing Agricultural University, Nanjing, 210095, P.R. China*
[2]*College of Pharmaceutical Sciences, Chengdu Medical College, Chengdu, 610500, P.R. China*
[*]For correspondence: gqs@njau.edu.cn; chenyuhang221@126.com
[†]Contributed equally to this work and are co-first authors

Received 30 August 2019; Accepted 28 September 2019; Published 16 January 2020

Abstract

Prunella vulgaris L. is a widely used traditional Chinese medicine containing a variety of secondary metabolites, but the molecular mechanism of the secondary metabolite synthesis pathway has not been determined. In this study the transcriptomes of roots, stems, leaves and flowers of *P. vulgaris* (seedling, bud stage and flowering stage) were sequenced using Illumina HiSeq 4000. *De novo* assembly was performed to generate a total of 146710 unigenes with an average length of 651 bp using Trinity software. Through blast alignment with 7 public databases, a total of 57825 unigenes annotated to non-redundant protein sequences (NR), 51101 unigenes annotated to nucleotide sequences (NT); 25528 unigenes annotated to Kyoto Encyclopedia of Genes and Genomes (KEGG), of which 25528 unigenes metabolic pathway-related genes. There were 52136 unigenes in cell components, biological processes and molecular functions in the Gene Ontology (GO). In addition, there were 58382, 51224, 33023 and 52136 unigenes annotated to a manually annotated and reviewed protein sequence database (Swiss-Prot), protein family (PFAM), the Clusters of Orthologous Groups of proteins (COG), and GO, respectively. DEGs were analyzed and enriched in GO and KEGG to predict their function. This study also identified 18830 SSRs. This analysis of the transcriptome of *P. vulgaris* may provide a basis for the study of the biosynthesis of secondary metabolites, discovery of functional genes, and development of molecular markers. © 2020 Friends Science Publishers

Keywords: *Prunella vulgaris*; Transcriptome; Functional information; DEGs; SSR

Introduction

Prunella vulgaris L. is a perennial herb belonging to the *Prunella* genus from *Lamiaceae*, known as “self-heal.” *P. vulgaris* is widely distributed in China (Liao *et al*. 2012), Korea (Han *et al*. 2009), Japan (Hisashi *et al*. 1990), Pakistan (Fazal *et al*. 2016), the Czech Republic (Škottová *et al*. 2004) and other Asian and European countries. In China, this herb is widely distributed throughout central, south-western and south-eastern areas, including Sichuan, Hubei, Jiangxi, Henan, Anhui, Jiangsu and Zhejiang provinces (Zhang *et al*. 2013). Moreover, *P. vulgaris* is mostly found in streams, wet grass, bushes, field edges or roadsides near the south of China (Bai *et al*. 2016). *P. vulgaris* has been used in ttraditional medicinal preparations for over 2000 years (Chen *et al*. 2019), and it is commonly used for its antifebrile and sedative properties and the treatment of detumescence, thyroid gland malfunction and mastitis. The dried spica of *P. vulgaris* is called a “Xiaku-cao” in Chinese, which means that it will wither after summer and is a standard medicinal material in the Chinese Pharmacopoeia. Phytochemical analysis showed that *P. vulgaris* contains a large amount of phenolic acid, triterpenes, flavonoids and polysaccharide (Wang *et al*. 2000; 2011; Cheung and Zhang 2008; Jiang *et al*. 2008). Previous studies have indicated that *P. vulgaris* exhibits a wide range of pharmacological bioactivities, including hepatoprotection, anti-viral, anti-hyperglycaemia, antitumor, anti-inflammatory, anticancer, anti-HSV, anti-HIV (Yao *et al*. 1992; Xu *et al*. 1999) and immune modulation activity (Feng *et al*. 2010; Li *et al*. 2015a). In addition to the application of this herb as a traditional Chinese medicine, the *P. vulgaris* plant is also used for ornamental purposes (as flowers), in functional tea (the air-dried plants) and as a medicinal leafy vegetables (Chen *et al*. 2012a–c; Li *et al*. 2015b).

Demand for *P. vulgaris* has grown steadily in recent years due to its medicinal and industrial importance. Although the current cultivation of *P. vulgaris* partially meets market demand, the problem of declining quality of medicinal materials has been unavoidable (Chen *et al*. 2013). In recent years, artificial culture and regeneration systems

To cite this paper: Li Z, Y Chen, Q Guo, C Wang, L Cao, Q Qin, M Zhao, C Li (2020). *De novo* sequencing and transcriptome analysis of *Prunella vulgaris* during development: a cross-databases comparison. *Intl J Agric Biol*, 23:302–310

have been well studied (Kour *et al*. 2014), and the chemical diversity and genetic variation of the population has been highlighted (Li *et al*. 2012; Xu et al. 2018). However, only approximately 1131 nucleotide sequences and 294 proteins were recorded in GenBank, which severely impedes the study of functional genes, thereby hindering the process of cultivating elite varieties by molecular means.

Illumina HiSeq is the next-generation sequencing technology platform with the advantage of being able to generate large numbers of sequences at lower cost and in less time (Mutz *et al*. 2013). Furthermore, transcriptome sequencing (RNA-Seq) can facilitate the discovery of new genes and further deepen the understanding of gene regulation and complex networks (Hua *et al*. 2011), and RNA-Seq overcomes previous expression microarray constraints (Martin and Wang 2011; Kudo *et al.* 2016), and provides a more complete view of the protein coding region (Li *et al.* 2012). In recent years, RNA-Seq studies on such model plant species such as *Oryza sativa* (Huang *et al*. 2019; Yang *et al*. 2015) and *Arabidopsis* (Ma *et al*. 2018) have indicated that this technique is well suited for investigating the complexity of transcription. Because RNA-Seq has no restrictions on existing transcript sequences, it is a strong impetus for the development of functional genomics of non-model organisms (Grabherr *et al*. 2011) and molecular markers for marker-assisted breeding (Liao *et al*. 2012). Therefore, many medicinal plants were sequenced from the transcriptome by RNA-Seq (Gao *et al*. 2014; Xiang *et al*. 2014). Although *Salvia miltiorrhiza*, which belongs to the *Lamiaceae* family, has been the subjected to in-depth genome sequencing, transcriptome analysis of *P. vulgaris* in different periods and in different parts has not been reported. In this study, we analyzed the transcriptome results of *P. vulgaris* by using the RNA-Seq technology on a Hiseq 4000 platform. Moreover, many simple sequence repeat (SSR) markers and genes related to secondary metabolite synthesis were discovered, which will contribute to the research of *P. vulgaris*-assisted breeding and compositional synthesis mechanisms.

Materials and Methods

Plant materials and RNA isolation

Seeds of *P. vulgaris* were sown at the experimental station of Chengdu Medical College, Chengdu, China (latitude: 30°49′17″ N, longitude: 104°17′49″ E, altitude: 493 m) on 21 October 2015. Samples were collected at the seedling stage (fresh roots, leaves and stems), squaring stage (fresh leaves, stems and flower buds) and full-flowering stage (fresh leaves, stems and spicas) from 7 April to 5 May 2016. Fresh tissue was washed with sterile water, wrapped in foil, immediately frozen in liquid nitrogen and stored at -80°C before use. Total RNA from these nine samples was extracted using TRIzol (Invitrogen Co., USA).

cDNA library construction and sequencing

The RNA-Seq library was constructed by the Novogene Company (Tianjin, China) as follows: Magnetic beads with oligonucleotides (dT) enrich eukaryotic mRNA. Afterwards, the fragmentation buffer was used to break the mRNA into short fragments. Single-stranded cDNA was synthesized by random hexamers using mRNA as a template, and then dNTP and DNA polymerase I and RNase H acted together to synthesize double-stranded cDNA in a buffer. Purification of double-stranded cDNA was performed using AMPure XP beads. The cDNA was first end-repaired, A-tailed and ligated to the sequencing linker, and the fragment size was selected using AMPure XP beads. Finally, the product was purified by PCR and purified using XP beads to obtain the best library.

RNA-seq data processing and transcriptome analysis

The raw reads contained connectors with trimming adapter and low quality reads (reads containing more than 50% bases with Q-value≤20). Cleaned and qualified reads were then assembled into unigenes using Trinity software (Grabherr *et al*. 2011). All the unigenes were compared with public databases, such as the GO, KOG, KEGG, NR, NT and Swiss-Prot databases. To explore the macroscopic distribution of gene function in species, we obtained the Gene Ontology (GO) annotation of Unigenes using the Blast2GO program (Götz *et al.* 2008) through WEGO software (Ye *et al.* 2006).

SSR detection and differential expression unigene analysis

The SSR sites were performed on all unigenes in the transcriptome using MISA (Version 1.0, default parameters; the minimum number of repetitions for each unit size is: 1–10, 2–6, 3–5, 4–5, 5–5, and 6–5), and the density distribution of gene transcripts was counted for different SSR types. Differential expression unigenes were analysed by standardizing the read count data using TMM, then analyzing it with DEGseq software, screening unigenes that met the threshold (q value<0.005 and |log2FoldChange|>1) and performing GO function and KEGG Pathway analysis.

Results

Transcriptome sequencing and *De novo* assembly

The cDNA libraries from four tissues (root, leaf, stem and spica) of *P. vulgaris* during the three developmental periods: vegetative root (VR) and vegetative stem (VS), vegetative leaf (VL), squaring stage stem (SST), squaring stage spicas (SSP), squaring stage leaf (SL), full-flowering stem (FST), full-flowering spicas (FSP) and full-flowering leaf (FL) generating approximately 42~53 million paired-

Table 1: Sequence yield and quality in transcriptome of *P. vulgaris*

Sample	Raw Reads	Clean Reads	Error (%)	Q20 (%)	Q30 (%)	GC (%)
V_R	42548090	41492962	0.01	98.14	95.31	47.18
V_S	50681256	49857146	0.01	98.21	95.55	46.61
V_L	48312908	47234206	0.01	97.79	94.60	48.18
S_S_T	45833726	45071082	0.01	98.21	95.54	46.63
S_S_P	49279706	48459508	0.01	98.20	95.53	46.38
S_L	43954828	42815074	0.01	97.37	93.47	47.93
F_S_T	53198386	52279650	0.01	98.01	95.02	46.48
F_S_P	47349998	46585006	0.01	98.20	95.54	46.34
F_L	45236316	43991976	0.01	97.34	93.47	47.94

V_R was expressed as vegetative root of *P. vulgaris*; V_S was expressed as vegetative stem of *P. vulgaris*; V_L was expressed as vegetative leaf of *P. vulgaris*; S_S_T was expressed as squaring stage stem of *P. vulgaris*; S_S_P was expressed as squaring stage stem of *P. vulgaris*; S_L was expressed as squaring stage leaf of *P. vulgaris*; F_S_T was expressed as full-flowering stem of *P. vulgaris*; F_S_P was expressed as full-flowering spicas of *P. vulgaris*; F_L was expressed as full-flowering leaf of *P. vulgaris*

Table 2: Length characteristics and frequency distribution of stitching length of *P. vulgaris*

Length statistics (bp)	Transcripts	Unigenes	Transcript length interval	Transcripts	Unigenes	Database	Unigenes
Min/max length	201/13317	201/13317	201–500 bp	115036	100967	NR/NT	57825/51101
Mean length	1015	651	501–1000 bp	37950	22531	KEGG/ PFAM	25528/ 51224
Median length	490	330	1001–2000 bp	37800	12996	Swiss-Prot	58382
N50/ N90	1981/358	1134/255	>2001 bp	36310	10216	COG/GO	33023/52136
Total	230426310	95525973	Total	227096	146710	Total	146710

N50/N90 was expressed as sorting splicing transcripts from long to short, lengthing the length of the transcript to a length of not less than 50%/90% of the total length of the spliced transcript. NR was expressed as non-redundant protein sequences; NT was expressed as nucleotide sequences; KEGG was expressed as Kyoto Encyclopedia of Genes and Genomes; Swiss-Prot was expressed as manually annotated and reviewed protein sequence database; PFAM was expressed as protein family; COG was expressed as the Clusters of Orthologous Groups of proteins; GO was expressed as the Gene Ontology

end reads and approximately 41~52 million clean reads (approximately 6.22~7.84 Gb) from each tissue. The percentages of Q20, Q30 and GC were 97.84~98.21%, 93.47~95.55% and 46.34~48.18%, respectively (Table 1). All high-quality clean reads were assembled into 146710 unigenes using Trinity, with a mean length of 651 bp, an N50 length of 1134 bp and a total length of 95525973 bp (Table 2). Eight percent (12996) of the unigenes ranged from 1001 to 2000 bp, and 10216 (6%) of the unigenes were longer than 2001 bp. Fifteen percent (22531) of the unigenes ranged from 501 to 1000 bp. Most of the unigenes (100967) ranged from 201 to 500 bp, accounting for 68.82% of the total unigenes (Table 2).

Sequence annotation and classification

All of the unigenes were searched against seven public databases for functional annotation. 57825 (39%) unigenes annotated within the NR, 51101 (34%) unigenes annotated within the NT, 25528 (17%) unigenes annotated within the KEGG, 58382 (39%) unigenes annotated within the Swiss-Prot, 51224 (34%) unigenes annotated within the PFAM, 33023 (22%) unigenes annotated within the COG, and 52136 (35%) unigenes annotated within the GO (Table 2).

NR classification

The majority (31.6%) unigenes matched proteins from *Sesamum indicum*, followed by those from *Erythranthe guttata* (9.2%), *Hordeum vulgare* (3.8%), *Guillardia theta* (1.8%), *Solanum lycopersicum* (1.7%) and other species (52.0%) in the NR database (Fig. 1). When the e-value was smaller than 1.0e-45，42.4% of the mapped unigenes had strong homologs, while 57.6% of unigenes had a similarity when the e-value varied from 1e−5 to 1.0e−45. A 42.5% of mapped Unigenes ranged from 60 to 80% with known sequences, 43.5% with similarity above this range, and 14.0% with less than 60%.

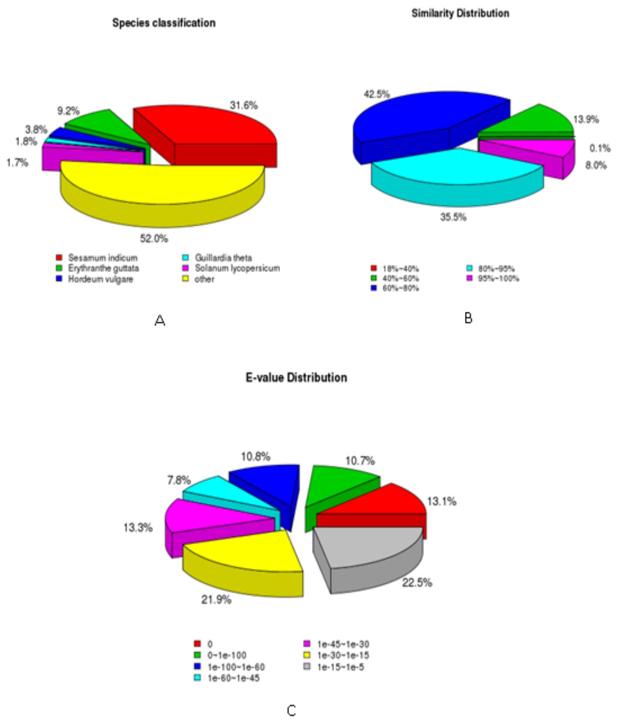

Fig. 1: Percent distribution and quality of unigenes assigned to different species in the NR database

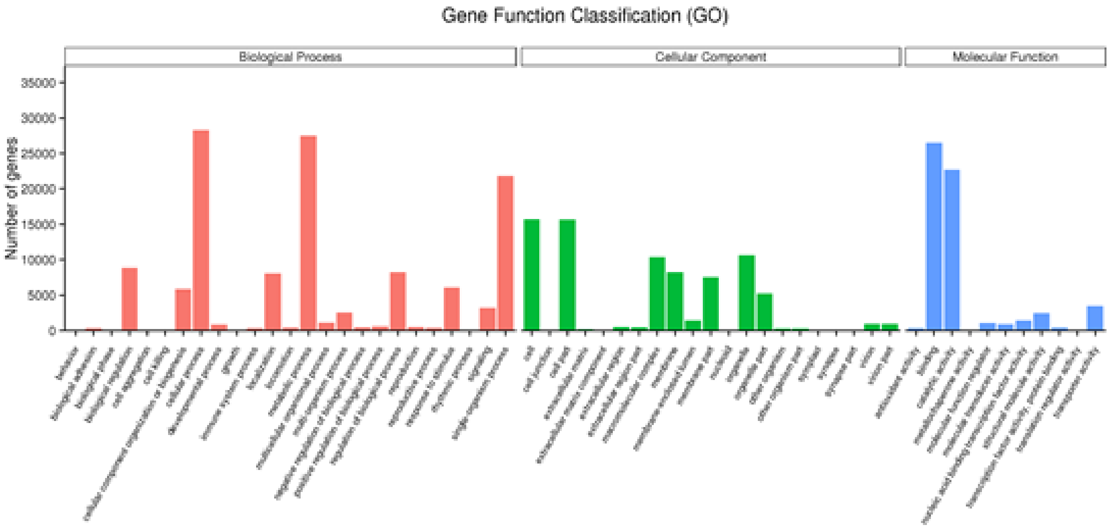

Fig. 2: Gene ontology classification of *P. vulgaris* transcriptome

Gene ontology classification

A total of 57825 unigenes of *P. vulgaris* were classified into three main GO categories. A total of 52136 unigenes (35.53%) were assigned 57 functional groups and allocated to three gene ontology (GO) categories at the second level, including 125774 unigenes of biological process, 59434 unigenes of molecular function and 78718 unigenes of cellular component (Fig. 2). Under the biological process category, cellular process (28289, 54.26%), metabolic process (27465, 52.68%), and single-organism process (21832, 41.88%) were prominently represented. For the molecular function category, binding (26523, 50.87%) and catalytic activity (22711, 43.56%) represented the majority of unique sequences. In the cellular component group, unique sequences related to cell (15721, 30.15%), cell part (15711, 30.14%), organelle (10651, 20.43%), and macromolecular complex (10368, 19.89%) were well-represented categories (Fig. 2).

Functional classification by KOG

Functional prediction and classification of all unigenes revealed that 33023 unigenes could be assigned to one or more of the 26 COG classification categories (Fig. 3). Among the 26 KOG categories, the largest cluster was predicted as general function (4611, 13.96%), followed by posttranslational modification, protein turnover and chaperones (4316, 13.07%), signal transduction mechanisms (2993, 9.06%), energy production and conversion (2530, 7.66%) intracellular trafficking, secretion, translation, ribosomal structure and biogenesis (3868, 11.71%), and vesicular transport (1935, 5.86%), amino acid transport and metabolism (1652, 5.00%), lipid transport and metabolism (1634, 4.95%), transcription (1616, 4.89%), RNA processing and modification (1589, 4.81%), carbohydrate transport and metabolism (1579, 4.78%), secondary metabolite biosynthesis, transport and catabolism (1375, 4.16%), cytoskeleton (1209, 3.66%) and function unknown (1060, 3.21%).

Functional classification by KEGG

KEGG founded in 1995 by The Kanehisa Laboratory of the Bioinformatics Center of Kyoto University, Japan, was established to provide researchers with more biological information. In this study, 1454, 745, 7125, 13529, and 911 unigenes were annotated into cellular processes, environmental information processing, genetic information processing, metabolism, and organic system pathways, respectively (Fig. 4). A total of 25528 unigenes of *P. vulgaris* have significant matches in the active biological pathways in the KEGG database, all with corresponding Enzyme Commission (EC) numbers and assigned to 131 KEGG pathways.

Genetic analysis of annotations into metabolic pathways revealed that there were 307 unigenes in the secondary metabolism that were annotated to the phenylpropanoid biosynthesis (ko00940) pathway, and 110 unigenes were annotated to the isoquinoline alkaloid biosynthesis (ko00950) pathway, 67 unigenes were annotated to the monocyclic lactam biosynthesis (ko00261) pathway, 56 unigenes were annotated to the flavonoid biosynthetic pathway (ko00941), 22 unigenes were annotated to the caffeine metabolism (ko00232) pathway, and 19 unigenes were annotated to thiopurine biosynthesis (ko00966) pathway, 18 unigenes were annotated to the biosynthesis of flavonoids and flavonols (ko00944) pathway, 17 unigenes were annotated to the anthocyanin biosynthesis (ko00942) pathway, 11 unigenes were annotated to the isoflavone biosynthesis (ko00943) pathway, and only 7, 3 and 1 unigenes were noted for the

Table 3: The number of unigenes was annotated to secondary metabolic pathway

KO number	Secondary metabolic pathway	Number of unigenes
ko00942	Anthocyanin biosynthesis	17
ko00965	Betalain biosynthesis	7
ko00232	Caffeine metabolism	22
ko00332	Carbapenem biosynthesis	1
ko00944	Flavone and flavonol biosynthesis	18
ko00941	Flavonoid biosynthesis	56
ko00966	Glucosinolate biosynthesis	19
ko00901	Indole alkaloid biosynthesis	3
ko00943	Isoflavonoid biosynthesis	11
ko00950	Isoquinoline alkaloid biosynthesis	110
ko00261	Monobactam biosynthesis	67
ko00940	Phenylpropanoid biosynthesis	307

Table 4: The analysis of differentially expressed unigenes

S v F	Up-regulated unigenes	Down-regulated unigenes	Total
Leaf	781	418	1199
Stem	584	579	1163
Spica	1261	803	2064

S v F was expressed as Squaring vs Flowering stage

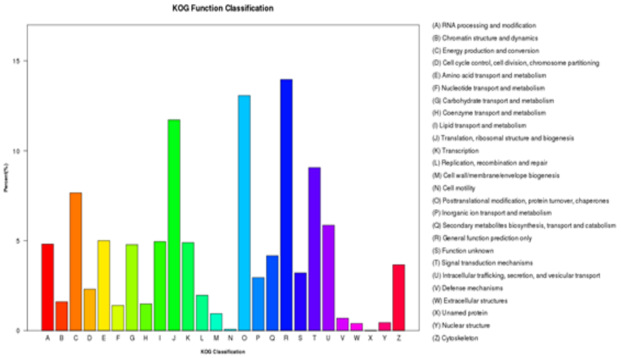

Fig. 3: Histogram presentation of clusters of euKaryotic Ortholog Groups (KOG) classification

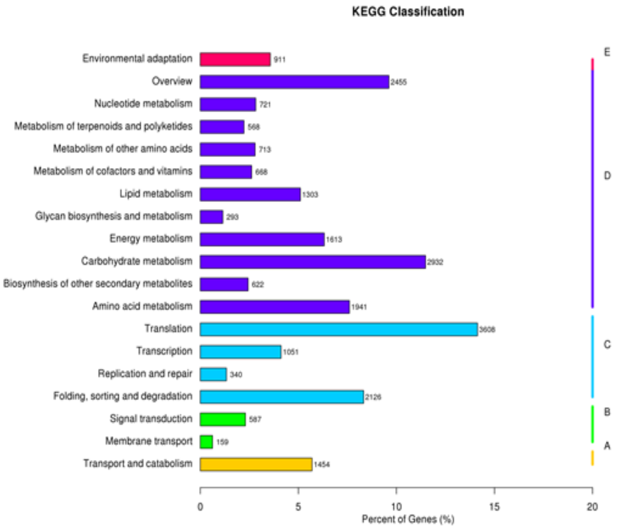

Fig. 4: KEGG pathway annotation of unigenes from *P. vulgaris*

beet pigment biosynthesis pathway (ko00965), indole alkaloid biosynthesis pathway (ko00901), and carbapenem biosynthesis pathway (ko00332), respectively (Table 3).

Analysis of differential expression of unigenes in *P. vulgaris*

Differential expression analysis was carried out in *P. vulgaris* transcripts by DGEseq software. Out of the total differentially expressed genes in roots, stems and leaves at the seedling stage, 3131 unigenes were upregulated and 3289 were downregulated; There were 2993 upregulated unigenes and 2769 downregulated unigenes in stems, leaves and spica at the squaring stage, while there were 2121 upregulated unigenes and 2478 downregulated unigenes in stems, leaves and spica during flowering stage. Even in the same part of *P. vulgaris*, there were some differences in gene expression levels at different growth stages (Table 4). Compared with the same part of *P. vulgaris* in the squaring stage and flowering stage, there were 1199 unigenes with different expression in the leaves, of which 781 were upregulated and 418 were downregulated; there were 2064 unigenes with different expression in the spica, of which 1261 were upregulated and 803 were downregulated. In addition, there were 1163 unigenes with different expression in the stem, of which 584 were upregulated and 579 were downregulated.

Functional classification of differentially expressed unigenes by GO

Differentially expressed unigenes in the transcripts of *P. vulgaris* were enriched into different GO functions (Table 5). A total of 925 differentially expressed unigenes were obtained in the leaves of *P. vulgaris* during the squaring stage and flowering stage, and 329 and 596 unigenes were upregulated and downregulated, respectively. The total differential unigenes had up to 542 enriched in catalytic activities (0003824), accounting for 58.59%, 210 downregulated unigenes were enriched in catalytic activity (0003824), accounting for 63.83%, and 234 unigenes were upregulated in the enrichment of ion binding molecular (0043167), accounting for 39.26%; also, 859 differential expression unigenes were obtained in the stems, and 429, 430 unigenes were upregulated and downregulated, respectively. The total differential unigenes and the downregulated unigenes were also the most enriched in catalytic activity, with 528 (61.47%) and 289 (67.21%), respectively, and 76 unigenes were upregulated in the enrichment of oxidoreductase activity molecular (0016491), accounting for 17.72%; 1530 differential expression unigenes were obtained in the stems, and 931, 599 unigenes were upregulated and downregulated, respectively. The total differential unigenes, down-regulated unigenes and upregulated unigenes were the most enriched in catalytic activity (0003824), with 918 (60%), 375 (62.6%) and 543 (58.32%), respectively.

Table 5: The analysis of differentially expressed genes enriched in GO / KEGG function

S v F	URU (GO/KEGG)	MF(GO/KEGG)	DRU (GO/KEGG)	MF(GO/KEGG)	TDEU (GO/KEGG)	MF(GO/KEGG)
Leaf	329/234	0043167(234)/ ko04626(29)	596/207	0003824(210)/ ko00710(18), ko00500(16)	925/441	0003824(542)/ ko04626(30)
Stem	429/180	0016491(76)/ ko00052(16)	430/207	0003824(289)/ ko00040(16), ko00520(15)	859/361	0003824(528)/ ko00500(23)
Spica	931/443	0003824(543)/ ko00500(41)	599/243	0003824(375)/ ko00940(18)	1530/361	0003824(918)/ ko00500(50)

S v F was expressed as Squaring vs flowering stage; URU was expressed as Up-regulated unigenes; DRU was expressed as Down-regulated unigenes; MF was expressed as Main function; TDEU was expressed as Total differentially expressed unigenes. 0043167 were expressed as ion binding molecular; 0016491 were expressed as oxidoreductase activity molecular; 0003824 was expressed as catalytic activity. ko04626 was expressed as plant-pathogen interaction; ko00052 was expressed as galactose metabolism; ko00500 was expressed as starch and sucrose metabolism; ko00710 was expressed as photosynthetic organisms; ko00040 was expressed as pentose and glucuronic acid interconversion; ko00520 was expressed as amino sugar and nucleotide sugar metabolism; ko00940 was expressed as phenylpropanoid biosynthesis

Table 6: Distribution of identified SSRs using the MISA software

Motif	Numbers of the repeated								
	5	6	7	8	9	10	>10	Total	(%)
Mononucleotide	-	-	-	-	-	3496	3198	6694	35.55
Dinucleotide	-	3139	1822	1234	695	286	85	7261	38.56
Trinucleotide	3032	1127	381	28	-	2	1	4571	24.28
Tetranucleotide	208	33	1	1	-	1	1	245	1.30
Pentanucleotide	26	4	2	1	1	-	-	34	0.18
Hexanucleotide	14	6	4	1	-	-	-	25	0.13
Total	3280	4309	2210	1265	696	3785	3258	18830	100
Percentage	17.42	22.88	11.74	6.72	3.70	20.10	17.44	100	

Functional classification of differentially expressed unigenes by KEGG

There were differences in the enrichment of differentially expressed unigenes of *P. vulgaris* in the KEGG pathway (Table 5). At the squaring and flowering stages, 441 differentially expressed unigenes (leaves) were enriched in the KEGG pathway. The upregulated and the downregulated unigenes were 234 and 207, respectively. The total and upregulated differential unigenes were mainly enriched in the plant-pathogen interaction (ko04626) pathway, with 30 and 29 being observed, respectively, and 18 differentially expressed unigenes were enriched in carbon fixation in photosynthetic organisms (ko00710) pathway. At the squaring and flowering stages, 361 differentially expressed unigenes (stems) were enriched in the KEGG pathway, and 180 upregulated unigenes and 181 downregulated unigenes were observed. Twenty-three of the total differential unigenes were enriched in the starch and sucrose metabolism (ko00500) pathway, and 16 of the differentially expressed unigenes were enriched in the galactose metabolism (ko00052) pathway. Downregulated differential unigenes were primarily enriched in the starch and sucrose metabolism (ko00500) pathway, pentose and glucuronic acid inter-conversion (ko00040) pathway, amino sugar and nucleotide sugar metabolism (ko00520) pathways, each of which was annotated with 16, 16 and 15 pathways, respectively. At the squaring and flowering stages, 686 differentially expressed unigenes (stems) were enriched in the KEGG pathway, and the upregulated unigenes and the downregulated unigenes were 443 and 243, respectively. The total differential unigenes were mainly enriched in the starch and sucrose metabolism (ko00500) pathway, with 50 being observed. The upregulated differential unigenes were mainly enriched in starch and sucrose metabolism (ko00500) pathways, with 41, and 18 downregulated differential unigenes were enriched in the phenylpropanoid biosynthesis (ko00940) pathway.

SSR discovery: distribution and frequencies

Unigenes were subjected to SSR detection using MISA software, and the density distribution of different types of SSRs in gene transcripts was statistically analyzed (Table 6). The number of repeats ranged from 5 to 24, in which the number of mononucleotide repeats and dinucleotide repeats were 6694 (35.55%) and 7261 (38.56%), respectively, the number of different repeat unit sizes were 4571 (24.28%) for trinucleotide, 245 (1.30%) for tetranucleotide, and 34 (0.18%) for pentanucleotide, and the hexanucleotide repeat had a minimum of 25 (0.13%).

Discussion

The spicas of *P. vulgaris* contains a variety of active ingredients, and has a variety of pharmacological effects, and the demand for this herb is high (Wang *et al*. 2011; Chen *et al*. 2013; Li *et al*. 2015a). Thus, research on its secondary metabolism and flowering mechanism needs to be strengthened, but it requires higher cost due to genome sequencing. Unigene functional annotation and classification can provide clues for studying metabolic pathways within cells and the biological behavior of genes (Tang *et al.* 2014). This study annotated the obtained unigene into seven authoritative databases. Nr includes the protein coding sequence of the GenBank gene, Protein Data Bank, SwissProt protein sequence and protein sequences.

PFAM (Protein family), the most comprehensive classification system for protein domain annotation, proteins are composed of one domain, and the protein sequence of each specific domain is somewhat conserved. Both KOG and COG are NCBI-based gene orthologous relationships, where COG is directed against prokaryotes and KOG is directed against eukaryotes. COG/KOG combines evolutionary relationships to classify homologous genes from different species into different orthologue clusters. Swiss-Prot collects protein sequences that have been collated and studied by experienced biologists. The KEGG system analyses the metabolic pathways of gene products and compounds in cells and the database of the function of these gene products. The KO (KEGG ortholog) system links the various KEGG annotation systems, and KEGG has established a complete KO annotation system to perform functional annotation of the genome or transcriptome of the newly sequenced species.

The homology analysis of *P. vulgaris* unigenes to the NR database revealed that most of them had high homology with sesame, which indicated that the homologous relationship between *P. vulgaris* and sesame was relatively closer. Among the 57525 unigenes annotated to GO, 28289 (54.26%) and 27465 (52.68%) belonged to cellular processes and metabolic processes, respectively. There were 33023 unigenes found corresponding to functional information in 25 COG classification categories, most of which were annotated to "general function" and "posttranslational modification, protein turnover and chaperones." In addition, 25528 unigenes were annotated in the KEGG database, with the most genes annotated in the metabolic pathway, a plurality of genes involved in the synthesis of medicinal components of *P. vulgaris*, including phenylpropanoids and flavonoids. These unigene excavations and studies provide the basis for the discovery of more metabolic pathways in the relevant metabolic pathways of *P. vulgaris* and the study of key enzyme candidate gene clones in related experiments. In addition, by digging deep into the differentially expressed unigenes of *P. vulgaris*, it was found that there were a large number of differentially expressed unigenes in different periods and different parts. Compared with the same part of *P. vulgaris* at the squaring stage and flowering stage, the total differentially expressed genes in leaves, stems and spica were mainly enriched in the catalytic activity of GO function (0003824); the total differentially expressed genes of leaves were mainly enriched in plant-pathogens in KEGG pathway (ko04626). Moreover, the total differentially expressed genes of stems and spica were mainly enriched in starch and sucrose metabolism in the KEGG pathway (ko00500). In recent years, research on the value of summer medicinal herbs has received increasing attention. These analyses will clarify the differential genes and regulatory mechanisms of *P. vulgaris* stems, leaves and spicas at different stages, and provide a basis for further research on *P. vulgaris*.

A total of 146,710 transcripts were obtained from this study. Due to the unique nature of the database, the number of unigenes annotated for each database is also different. Among unigenes, the number of unigenes annotated to Swiss-Prot is up to 58382 and the minimum number of annotated to KEGG is 25528. The differentially expressed unigenes in the same part of *P. vulgaris* are also very different in the GO and KEGG databases at the squaring stage and flowering stage. Among the leaves of *P. vulgaris*, 925 differentially expressed unigenes were annotated to the GO database by approximately 2 times of the KEGG database, and 596 of the upregulated unigenes of GO were 2.9 times of the annotation to KEGG, and 329 of the downregulated unigenes of GO were 1.4 times of the annotation to KEGG; among the leaves of *P. vulgaris*, whether it is the total differential unigenes (859), the upregulated expression gene (430) or the downregulated expression unigenes (429) annotated to GO is approximately twice as much as the annotation to KEGG; among the spicas of *P. vulgaris*, the total differential unigenes (1530) annotated to GO is approximately 4.2 times that of the annotation to KEGG, and the upregulated (599) and downregulated (913) expression unigenes of GO are approximately 2 times the annotation to KEGG. These differences are mainly related to the characteristics of the database. The basis of the GO database is a single GO term, which is a tree-like structure with redundancy. The GO term is a pure gene set that does not define the interrelationship of genes; KEGG not only has a gene set, but also defines the complex interrelationship between genes and metabolites. Differences in the database and annotation of differentially expressed genes will provide the basis for studies of functional genes in specific directions.

SSR is a simple sequence repeat, also known as microsatellite sequence (MS), or short tandem repeat (SRT), which is a kind of DNA fragment formed by multiple tandem repeats of 1 to 6 nucleotides. The length is generally shorter, below 200 bp, and is abundant in the plant genome. The frequency of occurrence in different plants is relatively large, but most of them are (AT)n. SSR molecular markers have the advantages of co-dominance, good reproducibility, rich polymorphism, and easy operation (Liu *et al.* 2014; Yuan *et al.* 2014). These markers have become the main means of plant variety identification, genetic diversity analysis, genetic map construction and molecular marker-assisted breeding (Mi *et al.* 2015; Yang *et al.* 2018). We identified 24346 SSRs, and statistical analysis of different types of SSR, the highest proportion of dinucleotide repeats appeared, and the higher frequency of the motifs were A / T, AG/CT, AT/AT, AC/GT, AAG/CTT. The results of SSR analysis of *P. vulgaris* can provide a data basis for the study of genomic difference analysis, molecular marker development and genetic map construction of *Prunella* L.

Conclusion

NGS and RNA-Seq technology was used to sequence *P. vulgaris*. After *de novo* assembly and sequence annotation, a total of 146710 unigenes were obtained, and 25528 were annotated into the KEGG database, with the highest number of genes annotated into the metabolic pathway. In addition, through the analysis of differentially expressed genes, differences in functional genes at different flowering stages were found. We also identified 24346 SSRs. The analysis of the transcriptome of *P. vulgaris* may provide a basis for the study of the formation of secondary metabolites, the discovery of functional genes, and the development of molecular markers.

Acknowledgment

This study was funded by the programs of the National Nature Science Foundation of China (31500263, 30772730 and 81072986), Basic Research Program of Sichuan Province of China (2019YJ0368) and the Sichuan Province Undergraduates Innovating Experimentation Project (201813705027).

References

Bai YB, BH Xia, WJ Xie, YM Zhou, JC Xie, HQ Li, DF Liao, LM Lin, C Li (2016). Phytochemistry and pharmacological activities of the genus *Prunella*. *Food Chem* 204:483–496

Chen YH, XR Zhang, QS Guo, LP Cao, Q Qin, C Li, M Zhao, WM Wang (2019). Plant morphology, physiological characteristics, accumulation of secondary metabolites and antioxidant activities of *Prunella vulgaris* L. under UV solar exclusion. *Biol Res* 52:17

Cheung HY, QF Zhang (2008). Enhanced analysis of triterpenes, flavonoids and phenolic compounds in *Prunella vulgaris* L. by capillary zone electrophoresis with the addition of running buffer modifiers. *J Chromatogr A* 1213:231–238

Chen YH, QS Guo, ZB Zhu, LX Zhang (2012a). Changes in bioactive components related to the harvest time from the spicas of *Prunella vulgaris*. *Pharm Biol* 50:1118–1122

Chen YH, QS Guo, ZB Zhu, LX Zhang, XM Zhang (2012b). Variation in concentrations of major bioactive compounds in *Prunella vulgaris* L. related to plant parts and phenological stages. *Biol Res* 45:171–175

Chen YH, ZB Zhu, QS Guo, LX Zhang, XL Dai (2012c). Comparative analysis of the essential oil of flowers, leaves and stems of *Prunella vulgaris* L. *J Essen Oil Bearing Plants* 15:662–666

Chen YH, MM Yu, ZB Zhu, LX Zhang, QS Guo (2013). Optimisation of potassium chloride nutrition for proper growth, physiological development and bioactive component production in *Prunella vulgaris* L. *PLoS One* 8:e66259

Fazal H, BH Abbasi, N Ahmad, SS Ali, F Akbar, F Kanwal (2016). Correlation of different spectral lights with biomass accumulation and production of antioxidant secondary metabolites in callus cultures of medicinally important *Prunella vulgaris* L. *J Photochem Photobiol B-Biol* 159:1–7

Feng L, XB Jia, F Shi, Y Chen (2010). Identification of two polysaccharides from *Prunella vulgaris* L. and evaluation on their anti-lung adenocarcinoma activity. *Intl J Mol Med* 15:5093–5103

Grabherr MG, BJ Haas, M Yassour, JZ Levin, DA Thompson, I Amit, X Adiconis, L Fan, R Raychowdhury, QD Zeng, ZH Chen, E Mauceli, N Hacohen, A Gnirke, N Rhind, F di Palma, BW Birren, C Nusbaum, K Lindblad-Toh, N Friedman, A Regev (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 29:644–652

Gao W, HX Sun, H Xiao, G Cui, ML Hillwig, A Jackson, X Wang, Y Shen, N Zhao, L Zhang, XJ Wang, RJ Peters, L Huang (2014). Combining metabolomics and transcriptomics to characterize tanshinone biosynthesis in *Salvia miltiorrhiza*. *BMC Genom* 15:1–14

Götz S, JM García-Gómez, J Terol, TD Williams, SH Nagaraj, MJ Nueda, M Robles, M Talón, J Dopazo, A Conesa (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucl Acids Res* 36:3420–3435

Hua WP, Y Zhang, J Song, LJ Zhao, ZZ Wang (2011). *De novo* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. *Genomics* 98:272–279

Huang YM, HQ Chen, JR Reinfelder, XY Liang, JC Yi (2019). A transcriptomic (RNA-seq) analysis of genes responsive to both cadmium and arsenic stress in rice root. *Sci Tot Environ* 666:445–460

Han EH, JH Choi, YP Hwang, HJ Park, CY Choi, YC Chung, JK Seo, HG Jeong (2009). Immunostimulatory activity of aqueous extract isolated from *Prunella vulgaris*. *Food Chem Toxicol* 47:62–69

Hisashi K, S Noriko, H Akiko, O Haruo (1990). Sterol glucosides from *Prunella vulgaris*. *Phytochemistry* 29:2351–2355

Jiang SJ, LZ Zhao, YG Yu, LD Duan, BX Tan (2008). Study on optimization of supercritical fluid extraction conditions of ursolic acid from *Prunella vulgaris* Linn. Leaves. *Food Sci* 29:294–297

Kudo T, Y Sasaki, S Terashima, N Matsuda-Imai (2016). Identification of reference genes for quantitative expression analysis using large-scale RNA-seq data of arabidopsis thaliana and model crop plants. *Genes Genet Syst* 91:111–125

Kour B, M Azhar, S Kaul, MK Dhar (2014). In vitro regeneration and mass multiplication of *Prunella vulgaris* L. *Natl Acad Sci Lett* 37:81–86

Li C, L You, X Fu, Q Huang, S Yu, RH Liu (2015a). Structural characterization and immunomodulatory activity of a new heteropolysaccharide from *Prunella vulgaris*. *Food Funct* 6:1557–1567

Li C, Q Huang, X Fua, XJ Yue, RH Liu, LJ You (2015b). Characterization, antioxidant and immunomodulatory activities of polysaccharides from *Prunella vulgaris* Linn. *Intl J Biol Macromol* 75:298–305

Li X, C Zhu, CT Yeh, W Wu, EM Takacs, KA Petsch (2012). Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genom Res* 22:2436–2444

Liao L, QS Guo, ZY Wang, L Liu, ZB Zhu (2012(. Genetic diversity analysis of *Prunella vulgaris* in China using ISSR and SRAP markers. *Biochem Syst Ecol* 45:209–217

Liu YL, WQ Wang, JL Hou, R Zhang, WX Yang, FB Liu, SL Wei (2014). The optimization and primary application for SSR-PCR reaction system of the *Astragalus*. *Lishizhen Med Mat Med Res* 25:2227

Mutz KO, A Heilkenbrinker, M Lönne, JG Walter, F Stahl (2013). Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol* 24:22–30

Ma HF, DY Shen, YR Wu, H Xu, DL Dou (2018). RNA-seq for comparative transcript profiling of *Phytophthora capsici* during its interaction with *Arabidopsis thaliana*. *Physiol Mol Plant Pathol* 102:193–199

Mi YN, SH Zhang, Y Cai, XJ Liang, C Jiang, Y Qin (2015). Genetic diversity and relationship analysis of the Eucommia ulmoides germplasm using simple sequence repeat (SSR) markers. *Lishizhen Med Materia Med Res* 26:2507

Martin JA, Z Wang (2011). Next-generation transcriptome assembly. *Nat Rev Genet* 12:671–682

Škottová N, L Kazdová, O Oliyarnyk, R Večeřa, L Sobolován, J Ulrichová (2004). Phenolics-rich extracts from *Silybum marianum* and *Prunella vulgaris* reduce a high-sucrose diet induced oxidative stress in hereditary hypertriglyceridemic rats. *Pharmacol Res* 2:123–130

Tang XQ, YH Xiao, TT Lv, FQ Wang, QH Zhu, TQ Zheng, J Yang (2014). High-Throughput sequencing and *de novo* assembly of the *Isatis indigotica* transcriptome. *PLoS One* 9:e102963

Wang ZJ, YY Zhao, B Wang, TM Ai, YY Chen (2000). Depsides from *Prunella vulgaris*. *Chin Chem Lett* 11:997–1000

Wang YX, JB Yin, QS Guo, YH Xiao (2011). Dynamic change of active component content in different parts of *Prunella vulgaris*. *Zhongguo Zhong Yao Zhi* 36:741–745

Xu HX, SH Lee, SH Lee, RL White, J Blay (1999). Isolation and characterization of an anti-HSV polysaccharide from *Prunella vulgaris*. *Antiviral Res* 44:43–54

Xu W, L Chen, P He, J Yang, C Xu, B Wang, Z Wang, H Yang, M Xie, S Yang, L Qiu, Y Wang (2018). analysis of *Cf-12* tomato transcriptome profile in response to *Cladosporium fulvum* infection with Hisat, StringTie and Ballgown. *Intl J Agric Biol* 20:2590–2598

Xiang L, Y Li, YJ Zhu, HM Luo, CF Li, XL Xu, C Sun, JY Song, LC Shi, L He, W Sun, SL Chen (2014). Transcriptome analysis of the *Ophiocordyceps sinensis* fruiting body reveals putative genes involved infruiting body development and cordycepin biosynthesis. *Genomics* 103:154–159

Yao XJ, AW Mark, AP Michael (1992). Mechanism of inhibition of HIV-1 infection *in vitro* by purified extract of *Prunella vulgaris*. *Virology* 187:56–62

Yang SY, DL Hao, ZZ Song, GZ Yang, YH Su (2015). RNA-Seq analysis of differentially expressed genes in rice under varied nitrogen supplies. *Gene* 555:305–317

Ye J, L Fang, H Zheng, Y Zhang, J Chen, Z Zhang, J Wang, S Li, R Li, L Bolund, J Wang (2006). WEGO: a web tool for plotting GO annotations. *Nucl Acids Res* 34:293–297

Yuan Y, P Long, C Jiang (2014). Development and characterization of simple sequence repeat (SSR) markers based on a full-length cDNA library of *Scutellaria baicalensis*. *Genomics* 105:61

Yang W, W Jiang, GY Zhong, NN Liu, Q Chen, XY Wang (2018). Development and application of SSR molecular markers in *Saxifraga*. *Zhongguo Zhong Yao Zhi* 43:2057

Zhang X, SP Nie, JA Li, JM Li, Lin, C Li, MY Xie (2013). Extraction and determination of content of polysaccharides in *Prunella Vulgaris* Linn. *Food Res Dev* 34:81–85