Pol. J. Environ. Stud. Vol. 31, No. 4 (2022), 3561-3571
DOI: *10.15244/pjoes/146557* ONLINE PUBLICATION DATE: 2022-05-17

Original Research

Analysis of High-Throughput Transcriptome Sequencing of *Orychophragmus violaceus* Seedlings

Hongtao Hang[1, 2]*

[1]School of Karst Science, Guizhou Normal University, Guiyang, 550001, P.R. China
[2]State Engineering Technology Institute for Karst Desertification Control, Guiyang, 550001, P.R. China

Received: 27 October 2021
Accepted: 9 February 2022

Abstract

In order to obtain the genetic basis of transcriptome data of *Orychophragmus violaceus* seedlings, the transcriptome of *Orychophragmus violaceus* was paired-end sequenced by Illumina Novaseq 6000 platform, a total of 59174171 clean reads (17.75 Gb clean bases) were obtained, and 110919 unigenes were obtained after assembly by *de novo*, with the longest and shortest length of 15030, 301 bp and an average length of 784 bp. The N50 was 947 bp and the N90 was 396 bp. These unigenes were compared among seven public databases including Non-redundant protein sequences (NR), Nucleotide (NT), Swiss-prot protein database (Swiss-Prot), Protein family (Pfam), Eu-karyotic ortholog groups (KOG), Gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG), as a result of 75369 (67.94%), 69004 (62.21%), 62258 (56.12%), 56068 (50.54%), 27796 (25.05%), 56066 (50.54%), 32897 (29.65%) unigenes were annotated respectively. These annotation results showed that *Orychophragmus violaceus* had most homologous sequences with 13610 unigenes with *Quercus suber*. The GO annotations showed that 56066 unigenes were annotated with 219038, which were divided into 3 categories and 43 functional groups. The KOG annotations showed that 27796 unigenes were annotated and grouped into 25 functional categories. The KEGG annotations showed that 32897 unigenes were involved in 34 types of metabolic pathways and 305 metabolic pathway branches. A total of 18118 SSR sites and 112584 CDS sequences were detected according to analyzing the coding sequences and microsatellite. Base on the high-throughput transcriptome sequencing of *Orychophragmus violaceus*, with a large number of functional genes are excavated, which provide certain basic data support for the subsequent development of bioinformatics analysis such as molecular markers and functional metabolic pathways.

Keywords: *Orychophragmus violaceus*, adaptable plant, high-throughput sequencing, function annotation, bioinformation analysis

*e-mail: hanghongtao86@126.com

Introduction

Selecting adaptable plant species is widely regarded as principle barrier for the ecological restoration of rocky desertification [1]. The need to reduce soil erosion and increase community diversity has increased the demand for adaptable plant species as pioneer plants [2]. Therefore, considerable attention on exploring and selecting the ideal adaptable plant species due to their environmental adaptability such as higher productive forces, better resistance to adversity [3-8]. Although, abound of plants have been studies and used for ecological restoration, the ideal adaptable plant species source for ecological restoration of rocky desertification should consider environmental, plant species characteristics aspects [9-13].

Orychophragmus violaceus L., is an annual or biennial herb, which belonging to cruciferous plant [14]. *Orychophragmus violaceus* is widely distributed in various region of China, owing to its strong environmental adaptability such as good drought and salt tolerance [15-17], commonly known as the February Orchid. *Orychophragmus violaceus* is a common wild vegetable in early spring in China that has large tender stems and highly nutritious leaves which are rich in vitamins C, β-carotene and various mineral compounds [18], especially with higher content of micronutrients like Fe and Zn at the germination stages which can supply the deficiency of these elements in the human body. The seeds of *Orychophragmus violaceus* are rich in oil content (up to 50%), with high oil content of nearly more 10% than that of *Brassica rape* [19]. The seed oil contains relatively high amounts of linoleic (more than 50%) and moderate concentration of linolenic that can decrease the levels of total cholesterol (TC), triglyceride (Tg) in the serum of *Orychophragmus violaceus* which can soften blood vessels and prevent clot formation, and supply the deficiency of these elements in the human body. Moreover, the oil quality of *Orychophragmus violaceus* is superior to that of the most commonly used cooking oil such as rapeseed oil, cottonseed oil, peanut oil, sesame oil [20]. *Orychophragmus violaceus* has a strong breeding ability, is widely applied as ornamental flowers and grass sources [21]. It has attracted much attention in gardening and greening, health product development, pharmaceutical raw materials, biodiesel raw materials as so on.

At present, extensive research on the nutrient composition, physiology and biochemistry of *Orychophragmus violaceus* have been carried out [15-19, 21], and molecular level such as carbonic anhydrase isoenzyme genes [16], peroxide genes [22], enolpyruvylshikimic acid-3-phosphate synthetase genes [23] and chalcone synthase genes [24] have also been cloned. However, these above studies on genes clone of *Orychophragmus violaceus* were carried out by homologous comparison of other higher plants such as *Arabidopsis*, Rape, Mustard and other related gene conserved regions for cloning and analysis. To a large extent, these results limit the systematic research on the expression of more key genes of *Orychophragmus violaceus*.

In recent years, high-throughput transcriptome sequencing technology has been widely used as gene expression analysis in organisms [25]. Based on this technology, the gene transcription information of the research object in a certain state can be quickly obtained, as a result of abundant important functional genes being excavated and molecular mechanism on the differential biological traits being revealed. In this study, the transcriptome of *Orychophragmus violaceus* seedlings were paired-end sequenced using the high throughput sequencing technology (Illumina Novaseq 6000 platform). Functional annotation, classification and metabolic pathway analysis of unigenes were performed using bioinformatics methods. These results can provide scientific basis on mining molecular markers and relate functional metabolic pathways of *Orychophragmus violaceus* for further exploring.

Experimental

Biological Material

The test material, *Orychophragmus violaceus* L. seeds, was collected from Guiyang, Guizhou, China (26.35°N, 106.42°E). This experiment was carried out using Petri dish culture at the State Engineering Technology Institute for Karst Desertification Control. *Orychophragmus violaceus* seeds were washed with 75% ethanol, after three washes in sterile deionized water, 50 healthy seeds were randomly placed in 10-cm Petri dishes covered with two layers of moistened filter paper at 25ºC under darkness. After 20 days, three seedlings were randomly selected and collected, the surfaces were washed with sterile deionized water, and were placed in liquid nitrogen to rapidly freeze for reserve.

Total RNA Extraction, Library Construction and Sequencing

Total RNA was extracted using Plant RNA Reagent method (Invitrogen) [26] from the germinating seeds at 20 days. The concentration and integrity of RNA was determined by Qubit 2.0 fluorometer (Life Technologies) and Agilent 2100 bioanalyzer (Agilent Technologies). mRNA samples with polyA tail was enriched by Oligo (dT) magnetic beads, and broken into short fragments according to adding fragmentation buffer, then these short mRNA fragments were as template to synthesize the first strand cDNA with six base random primers, and to synthesize second strand cDNA with dNTPs base on polymerase I, the double-strand cDNA were purified by AMPure XP beads. The purified double-stranded cDNA was repaired and ligated with a sequencing adapter, various size

of fragment was selected by AMPure XP beads. Finally, the cDNA library was constructed by PCR enrichment. The transcriptome was paired-sequenced with 150 bp using the automated DNA sequencer (Illumina Novaseq 6000 sequencing platform- Illumina).

Cleaning and Assembly of Transcriptome Data and Unigene Annotation

The raw reads obtained by transcriptome sequencing was filtered with adaptors and low-quality sequences, as a result of obtaining high quality clean reads. And these short clean reads lacking a 100 bp clean region were spliced, clustered and *de novo* assembly to obtain non-redundant reads (unigenes) by Trinity software (version 2.5.1) based on the paired-end splicing method [27]. After clustering and assembly, the similarities between the unigenes and sequences deposited in public databases were detected by Basic Local Alignment Search Tool (BLAST) [28]. Therefore, these NCBI databases compose of NR, NT, Pfam, KOG, Swiss-Prot, KEGG, GO were using for BLAST searches and unigenes annotation.

Coding Sequences (CDS) and Simple Sequence Repeats (SSR) Site Analysis

The putative coding sequences and translations were searched with NR protein library and Swiss-Prot protein library in order of priority [29]. If the comparison is done, the open reading frame (ORF) coding information of the transcript will be extracted from the comparison result, and the coding region will be translated into amino acid sequence in the order of 5‘->3‘; if the comparison is unsuccessful, those unaligned sequences will be predicted its ORF by Estscan (version 3.0.3) [30]. SSR site of unigenes were detected and analyzed by the MISA software (version 1.0, default parameters; the minimum number of repetitions corresponding to each unit size are: 1-10, 2-6, 3-5, 4-5, 5-5, 6-5) [31].

Results and Discussion

Transcriptome Data Quality Analysis

Three samples of *Orychophragmus violaceus* seedlings were transcriptomic sequenced. 19208693, 22809277, 19601824 raw reads were obtained, respectively. After low-quality reads were discarded, 18422940, 21977529 and 18773702 clean reads were obtained, respectively. The base percentage of Q20 was between 97.54% and 98.23%, the base percentage of Q30 was between 93.54% and 94.71%, and the percentage of GC was between 43.99% and 46.66% (Table 1). The sequencing results of Q30 (94.20% on average) and N50 (947 bp) indicated that the high quality of transcriptome sequencing data was very reliable can be processed and analyzed in subsequent splicing and assembly.

Splicing and Assembly of Transcriptome Data

The full transcript was obtained by *de novo* splicing and assembly from the clean reads. This experiment obtained 277427 transcripts longer than 300 bp, with a total sequence length of 241655443 bp, N50 of 1106 bp, N90 of 420 bp, and an average length of 871 bp of *Orychophragmus violaceus* seedlings. Based on Trinity software, the transcript was determined as the longest sequence at each site with a total sequence of 110919 bp, total sequence length 87004315 bp, N50 was 947 bp, N90 was 396 bp. There were 65017 unigenes larger than 500 bp, 23035 unigenes larger than 1000 bp, the maximum unigene length was 15030 bp, the minimum unigene length was 301 bp, and the average length was 784 bp (Table 2, Fig. 1).

Gene Function Annotation and Metabolism Pathway Analysis

Comparing the unigenes of *Orychophragmus violaceus* with the NR, NT, Swiss-Prot, Pfam, KOG, GO and KEGG databases, we took the similarity greater

Table 1. Transcriptome sequencing quality analysis.

Sample	Raw Reads	Clean Reads	Error (%)	Q20 (%)	Q30 (%)	GC Content (%)
1	19208693	18422940	0.03	97.54	93.54	46.35
2	22809277	21977529	0.02	98.06	94.34	46.66
3	19601824	18773702	0.02	98.23	94.71	43.99

Table 2. The splicing assembly indexes of transcripts and unigenes.

Type	Total	Total basees	The longest transcript length	The shortest transcript length	Average length	N50	N90
Transcript	277427	241655443	15030	301	871	1106	420
Unigene	110919	87004315	15030	301	784	947	396

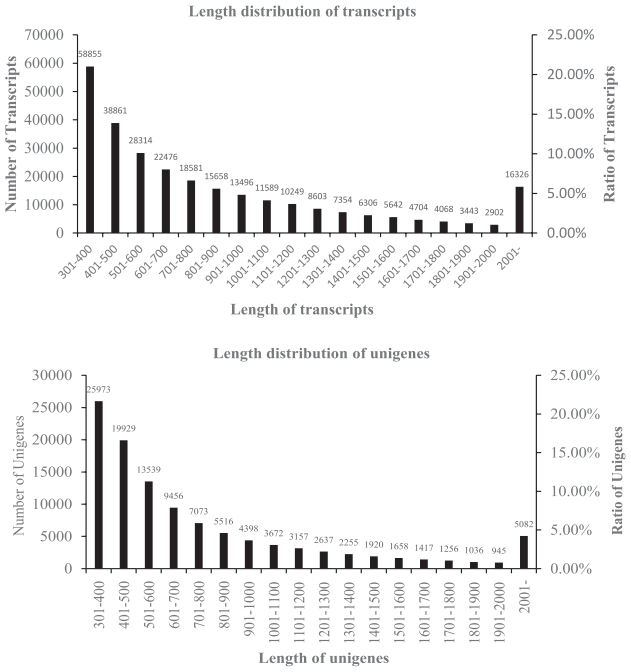

Fig. 1. Length distribution of transcripts and unigenes.

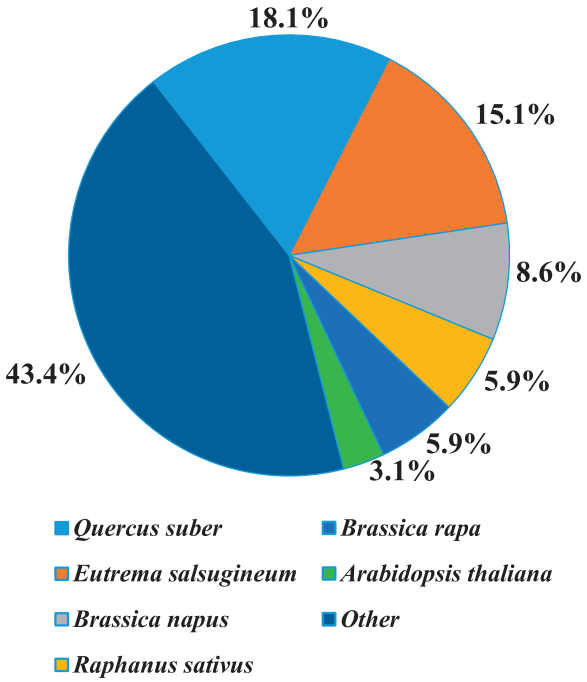

Fig. 2. Species distribution pie chart in NR database.

than 30%, and the annotations of e value less than 1e^{-5}, and merge all the annotation details. The results (see in Table 3) found that among 110919 unigenes of *Orychophragmus violaceus* seedlings, 75369 (67.94%) were annotated in NR database, 69004 (62.21%) were annotated in NT database, and 62258 (56.12%) were annotated in Swiss-Prot database, 56068 (50.54%) were annotated in Pfam database, 27796 (25.05%) were annotated in KOG database, 56066 (50.54%) were annotated in GO database, 32897 (29.65%) were annotated in KEGG database. The results that these 110919 unigenes of *Orychophragmus violaceus* were obtained by BLAST tool revealed that the inability of a large number of unigenes to reveal matching protein sequences in the NR and Swiss-Prot databases that is related to factors such as short unigene fragments, lack of gene annotation information in related databases, and the existence of new genes. The above data can provide important guidance information for the next step of the research on the domain of *Orychophragmus violaceus*.

NR Annotation

75369 unigenes from *Orychophragmus violaceus* were annotated in the NR database (see in Table 3). There were 13610, 11414, 6452, 4433, 4424 and 2308 unigenes of *Orychophragmus violaceus* similar to *Quercus suber*, *Eutrema salsugineum*, *Brassica napus*, *Raphanus sativus*, *Brassica rapa*, *Arabidopsis thaliana*, respectively. The ratio which accounted for 18.1%, 15.1%, 8.6%, 5.9%, 5.9%, 3.1% of the total number of unigenes annotated in the NR database, respectively. The remaining 43.4% of unigenes were annotated in 617 species (Fig. 2).

GO Annotation

GO database is an internationally standardized gene function classification database [32], which is used to comprehensively describe the biological characteristics of genes in different organisms. The unigenes of *Orychophragmus violaceus* were performed functional classification on gene biological characteristics base on GO database.

The results showed (Table 3, Fig. 3) that 56066 out of 110919 unigenes in *Orychophragmus violaceus* were functionally annotated in GO database, with an average of 3.91 GO annotations per transcript sequence. These annotated unigenes were divided into 3 categories 43 functional groups which compose of cellular

Table 3. Unigenes of *Orychophragmus violaceus* annotated proportion statistics in each database.

Database	NR	NT	Swiss-Prot	Pfam	KOG	GO	KEGG	Total unigenes
Number of comparisons	75369	69004	62258	56068	27796	56066	32897	110919
Ratio (%)	67.94	62.21	56.12	50.54	25.05	50.54	29.65	100

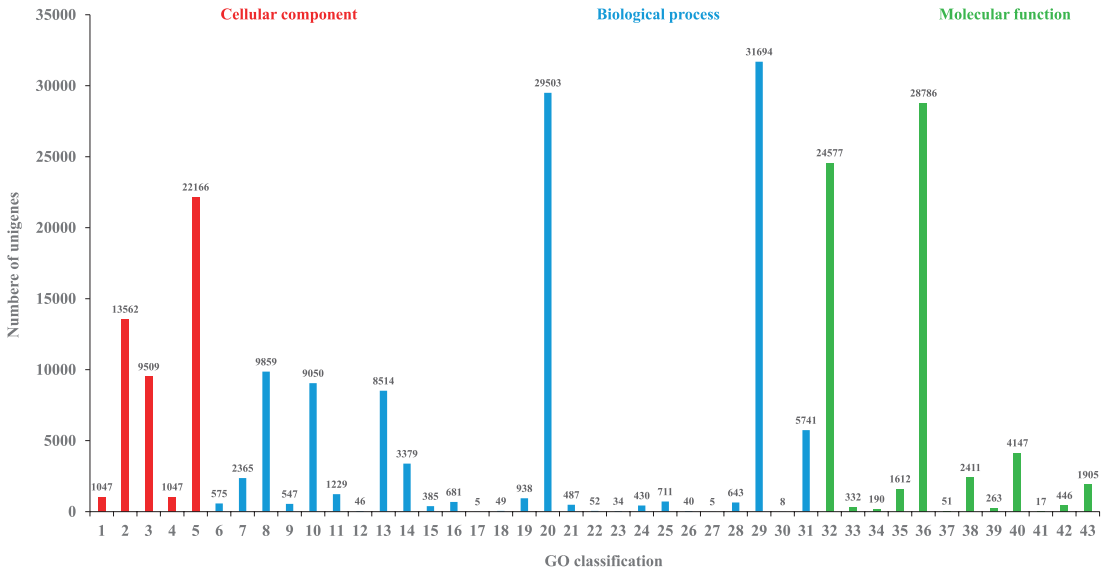

Fig. 3. GO annotation of unigenes of *Orychophragmus violaceus*.
Note: 1 Virion part; 2 Intracellular; 3 Protein-containing complex; 4 Virion; 5 Cellular anatomical entity; 6 Multi-organism process; 7 Interspecies interaction between organisms; 8 Biological regulation; 9 Positive regulation of biological process; 10 Regulation of biological process; 11 Multicellular organismal process; 12 Detoxification; 13 Localization; 14 Signaling; 15 Immune system process; 16 Reproduction; 17 Pigmentation; 18 Behavior; 19 Developmental process; 20 Metabolic process; 21 Biological adhesion; 22 Growth; 23 Intraspecies interaction between organisms; 24 Locomotion; 25 Negative regulation of biological process; 26 Rhythmic process; 27 Nitrogen utilization; 28 Reproductive process; 29 Cellular process; 30 Biomineralization; 31 Response to stimulus; 32 Catalytic activity; 33 Molecular transducer activity; 34 Small molecule sensor activity; 35 Molecular function regulator; 36 Binding; 37 Molecular carrier activity; 38 Structural molecule activity; 39 Antioxidant activity; 40 Transporter activity; 41 Cargo receptor activity; 42 Translation regulator activity; 43 Transcription regulator activity

components, molecular functions, and biological processes. Moreover, further analysis found that 47331 GO entries belong to 5 functional groups in the cell components, cellular anatomical entity account for the highest ratio with a value of 22166. 64737 GO items belong to 12 functional groups in molecular functions, and binding (28786 items) and catalytic activity (24577 items) account for a high proportion. 106970 GO items belong to 26 functional groups involved in biological processes, However, cellular processes (31694 items) and metabolic processes (29503 items) account for a higher proportion.

KOG Functional Prediction

Orychophragmus violaceus unigene sequences was KOG classified, as a result of a total of 27796 unigene sequences KOG functional annotations were obtained, involving 25 functional categories (Table 3, Fig. 4). Among them, Posttranslational modification, protein turnover, chaperones have the most transcripts with a value of 3773, accounting for 13.57%. Translation, ribosomal structure and biogenesis and general function prediction only are followed, the transcripts are 3653 and 3454, accounting for 13.14% and 12.43% respectively. The transcripts of extracellular structures and cell motility are only 39 and 30, accounting for 0.14% and 0.11%, respectively.

KEGG Metabolic Pathway Analysis

KEGG is a database that integrates genome, chemistry, and system function information [33]. It is a database that systematically analyzes the metabolic pathways of gene products in cells and the function of gene products. Comparing the *Orychophragmus violaceus* unigenes to the KEGG database, these results found that 32897 unigenes were participated in 34305 KEGG pathways branch, which were divided into 5 categories (see in Fig. 5) compose of cellular processes (A), environmental information processing (B), genetic information processing (C), metabolism (D), organic systems (E). Limitation of the length of this article, only annotated genes that account for >1% were listed (Table 4). Translation is the most amounts of annotated unigenes with a value of 4132, belong to genetic information processing branch. Secondly, signal transduction is environmental information branch with 3957 entries. The least annotated information which are only 14 entries is the signal molecule in the environmental information branch. These unigenes were mainly involved in ribosome, carbon metabolism, amino acid biosynthesis, protein processing in the endoplasmic reticulum, transcription, translation and other metabolic pathways.

Orychophragmus violaceus is an excellent oil plant for medicine [34] and food [35]. It is widely distributed throughout the country and has strong barren tolerance

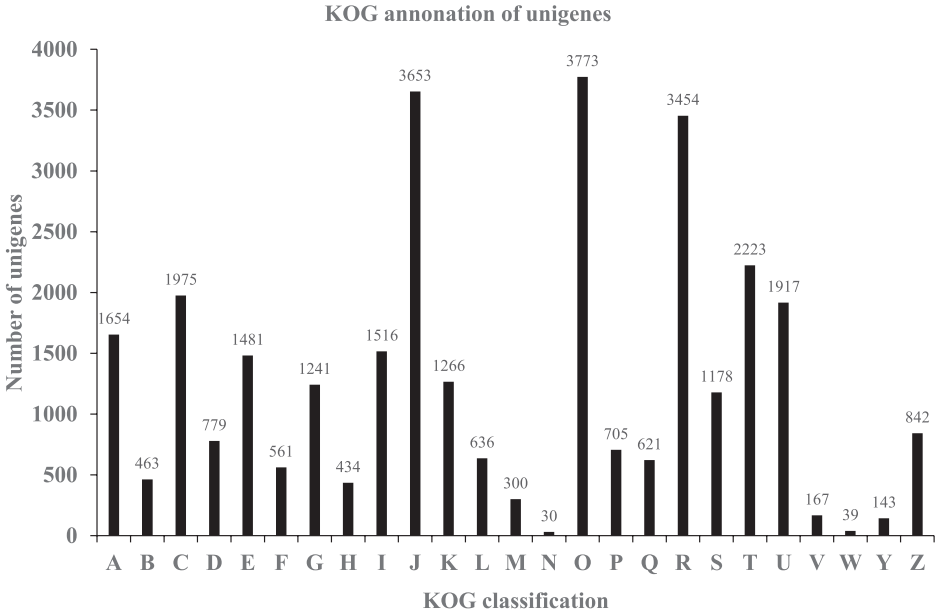

Fig. 4. KOG annotation of unigenes.
Note: A: RNA processing and modification; B: Chromatin structure and dynamics; C: Energy production and conversion; D: Cell cycle control, cell division, chromosome partitioning; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; J: Translation, ribosomal structure and biogenesis; K: Transcription; L: Replication, recombination and repair; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Posttranslational modification, protein turnover, chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Function unknown; T: Signal transduction mechanisms; U: Intracellular trafficking, secretion, and vesicular transport; V: Defense mechanisms; W: Extracellular structures; Y: Nuclear structure; Z: Cytoskeleton

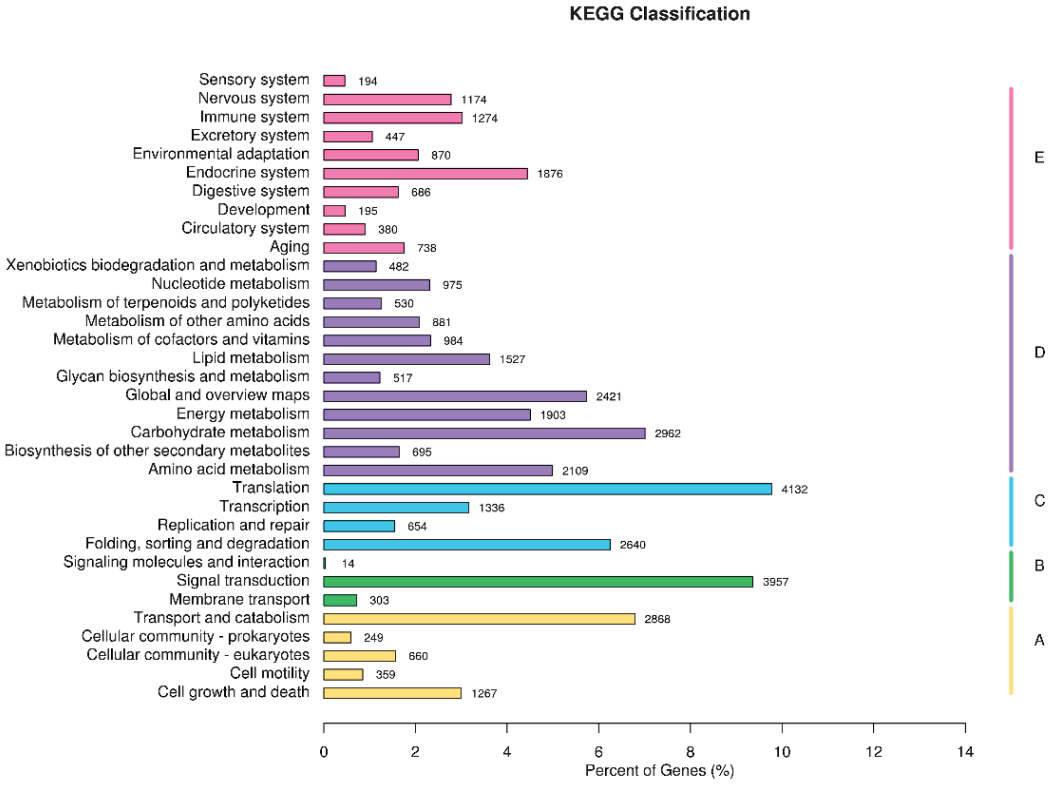

Fig. 5. KEGG pathway analysis of unigenes.
Note: A represents cellular processes, B represents environmental information processing, C represents genetic information processing, D represents metabolism, E represents organismal systems

Table 4. KEGG metabolic pathway of unigenes of *Orychophragmus violaceus* (annotated unigenes accounted for >1%).

Metabolic pathway	Metabolic pathway name	Number of unigenes
ko03010	Ribosome	2044
ko01200	Carbon metabolism	1457
ko01230	Biosynthesis of amino acids	1272
ko04141	Protein processing in endoplasmic reticulum	1125
ko03040	Spliceosome	1013
ko03013	RNA transport	869
ko04144	Endocytosis	855
ko00190	Oxidative phosphorylation	781
ko00230	Purine metabolism	756
ko04151	PI3K-Akt signaling pathway	654
ko04152	AMPK signaling pathway	619
ko03015	mRNA surveillance pathway	577
ko00010	Glycolysis / Gluconeogenesis	560
ko04075	Plant hormone signal transduction	559
ko04626	Plant-pathogen interaction	553
ko04910	Insulin signaling pathway	549
ko00240	Pyrimidine metabolism	535
ko04120	Ubiquitin mediated proteolysis	532
ko03008	Ribosome biogenesis in eukaryotes	530
ko04145	Phagosome	520
ko04110	Cell cycle	518
ko04016	MAPK signaling pathway - plant	486
ko04146	Peroxisome	475
ko00270	Cysteine and methionine metabolism	473
ko04114	Oocyte meiosis	472
ko00020	Citrate cycle (TCA cycle)	470
ko04142	Lysosome	469
ko03018	RNA degradation	467
ko04140	Autophagy - animal	461
ko00620	Pyruvate metabolism	454
ko00520	Amino sugar and nucleotide sugar metabolism	452
ko04138	Autophagy - yeast	438
ko04213	Longevity regulating pathway - multiple species	436
ko04150	mTOR signaling pathway	434
ko00630	Glyoxylate and dicarboxylate metabolism	432
ko04111	Cell cycle - yeast	429
ko00500	Starch and sucrose metabolism	423
ko04530	Tight junction	406
ko04071	Sphingolipid signaling pathway	399
ko04922	Glucagon signaling pathway	399
ko01212	Fatty acid metabolism	390
ko04721	Synaptic vesicle cycle	388
ko00564	Glycerophospholipid metabolism	381
ko01210	2-Oxocarboxylic acid metabolism	380
ko04024	cAMP signaling pathway	379
ko00250	Alanine, aspartate and glutamate metabolism	374
ko00970	Aminoacyl-tRNA biosynthesis	366
ko00480	Glutathione metabolism	363
ko00260	Glycine, serine and threonine metabolism	361
ko04068	FoxO signaling pathway	361
ko04810	Regulation of actin cytoskeleton	359
ko03050	Proteasome	358
ko04722	Neurotrophin signaling pathway	356
ko04921	Oxytocin signaling pathway	346
ko00280	Valine, leucine and isoleucine degradation	340
ko04666	Fc gamma R-mediated phagocytosis	340
ko04070	Phosphatidylinositol signaling system	336
ko04211	Longevity regulating pathway	330
ko04371	Apelin signaling pathway	329

[10, 15-17]. It is especially an important herb for ecological restoration as a suitable plant in karst areas [36, 37]. These data revealed these genes related to special metabolic pathways can provide basic data for further research.

SSR Analysis

Using MISA software to perform SSR analysis on unigene sequences, a total of 18118 SSR sites were detected. SSR types include single nucleotide to hexanucleotide repeat types (Table 5). Among them, the number of single nucleotides repeat types

Table 5. SSR site of unigenes analysis.

SSR types	Number of motif types	Motif types	Frequency
Single nucleotide	2	A/T	10537
		G/C	187
Di-nucleotide	4	AC/GT	424
		AG/CT	1991
		AT/AT	619
		CG/CG	13
Tri-nucleotide	10	AAC/GTT	689
		AAG/CTT	1349
		AAT/ATT	219
		ACC/GGT	346
		ACG/CGT	121
		ACT/AGT	84
		AGC/CTG	276
		AGG/CCT	317
		ATC/ATG	646
		CCG/CGG	122
Four-nucleotide	22	AAAC/GTTT	16
		AAAG/CTTT	14
		AAAT/ATTT	10
		AACC/GGTT	6
		AACG/CGTT	1
		AACT/AGTT	2
		AAGC/CTTG	1
		AAGG/CCTT	6
		AATC/ATTG	7
		AATG/ATTC	3
		ACAT/ATGT	7
		ACCG/CGGT	2
		ACCT/AGGT	5
		ACGG/CCGT	2
		ACTC/AGTG	2
		ACTG/AGTC	2
		AGAT/ATCT	5
		AGCC/CTGG	2
		AGCG/CGCT	2
		AGGC/CCTG	1
		ATCC/ATGG	5
		ATGC/ATGC	2
Five-nucleotide	22	AAAAC/GTTTT	2
		AAAAG/CTTTT	4
		AAAAT/ATTTT	4
		AAACC/GGTTT	3
		AAAGC/CTTTG	1
		AACAC/GTGTT	1
		AACCT/AGGTT	1
		AACGC/CGTTG	1
		AACTC/AGTTG	1
		AAGAG/CTCTT	1
		AATCG/ATTCG	1
		AATGC/ATTGC	1
		AATGG/ATTCC	2
		ACACC/GGTGT	1
		ACAGG/CCTGT	1
		ACCAG/CTGGT	1
		ACTCC/AGTGG	1
		AGAGG/CCTCT	2
		AGATC/ATCTG	1
		AGCAT/ATGCT	1
		ATATC/ATATG	1
		ATGCC/ATGGC	1
Hexa-nucleotide	33	AAAAAC/GTTTTT	1
		AAAACC/GGTTTT	2
		AAACAC/GTGTTT	1
		AAACAG/CTGTTT	1
		AAAGAG/CTCTTT	2
		AAAGCC/CTTTGG	1
		AAAGGG/CCCTTT	1
		AACAGC/CTGTTG	1
		AACCAT/ATGGTT	1
		AACCCT/AGGGTT	1
		AACCTC/AGGTTG	1
		AACTAC/AGTTGT	1
		AAGATC/ATCTTG	2
		AAGATG/ATCTTC	2
		AAGCAG/CTGCTT	2
		AAGCTC/AGCTTG	1
		AAGGAG/CCTTCT	3
		AATCTC/AGATTG	1
		ACACGG/CCGTGT	1
		ACCATC/ATGGTG	1
		ACCGCC/CGGTGG	1
		ACCTCC/AGGTGG	1
		ACGTCC/ACGTGG	1

Table 5. Continued.

Hexa-nucleotide	33	ACTCCC/AGTGGG	1
		ACTCTC/AGAGTG	1
		AGAGCT/AGCTCT	1
		AGAGGG/CCCTCT	1
		AGATCC/ATCTGG	2
		AGATGG/ATCTCC	1
		AGCCTC/AGGCTG	1
		AGCGGG/CCCGCT	1
		ATCGGC/ATGCCG	1
		ATGCCC/ATGGGC	2

is the largest, as many as 10724 and the ratio is 59.19%. There are 3047 di-nucleotide repeats, accounting for 16.82%. There are 4169 tri-nucleotide repeats, accounting for 23.01%. There are 103 four-nucleotide repeats, accounting for 0.57%. There are 33 five-nucleotide repeats, accounting for 0.18%. There are 42 hexa-nucleotide repeats, accounting for 0.23%. Among single-nucleotide repeats, A/T is the most, with 10537. Among di-nucleotide repeats, AG/CT is the most with 1991. Among three-nucleotide repeats, AAG/CTT is the most with 1349. Among the four-nucleotide repeats, AAAC/GTTT is the most, with 16. Among the five-nucleotide repeats, AAAAG/CTTTT and AAAAT/ATTTT are more numerous, each with 4. The six-nucleotide repeats are AAGGAG/CCTTCT at most as 3.

CDS of Unigenes Prediction

According to the priority order of the NR and Swiss-Prot databases, unigenes was aligned to the above two major protein databases. A total of 55963 CDS sequences were aligned, and 56621 CDS sequences were predicted by Estscan software (see in Fig. 6). Among the CDS sequences compared by BLAST tool, 82.09% are below 1000 bp, 17.90% are between 1 000-10 000 bp, and there are only 2 CDS sequences above 10000 bp. Among the CDS predicted by Estscan, 91.88% are below 1000 bp, 8.12% are between 1000 and 10000 bp, and there are only 3 CDS above 10000 bp.

Conclusions

In this study, the Illumina Novaseq 6000 sequencing technology was used for sequencing the transcriptome of the 20-day seedlings of *Orychophragmus violaceus* seedlings for the first time. The sequencing results of Q30 (94.20% on average) and N50 (947 bp) showed that the sequencing quality was very reliable and fulfilled the requirements of transcriptome analysis. The sequencing results were assembled to obtain 110919 unigenes compared to the NR, NT, Swiss-Prot, Pfam, KOG, GO and KEGG databases, a total of 18118 SSR sites were detected. Moreover, 55963 CDS sequences were predicted, and 56621 CDS sequences were predicted by Estscan tool. According to the GO database, 56066 unigenes annotated in *Orychophragmus violaceus* can be divided into 3 categories, 43 functional groups, and a total of 219038 GO entries. According to the KOG database, we have annotated the orthologous functions of the unigenes of *Orychophragmus violaceus*, and obtained 27796 unigenes which were divided into 25 functional categories in the eukaryotic functional system. The biological function of the open reading frame is unknown. KEGG functional annotations to 32897 unigenes were involved 34 branches of 305 metabolic pathways, mainly involved in ribosome, carbon metabolism, amino acid biosynthesis, protein processing in the endoplasmic reticulum, transcription, translation and other metabolic pathways. These findings revealed these genes related to special metabolic pathways can provide basic data for the subsequent research on the functional gene cloning and molecular marker development of *Orychophragmus violaceus*. More importantly, we plan to explore the key regulatory genes and related metabolic process of *Orychophramus violaceus* under abiotic stress environment.

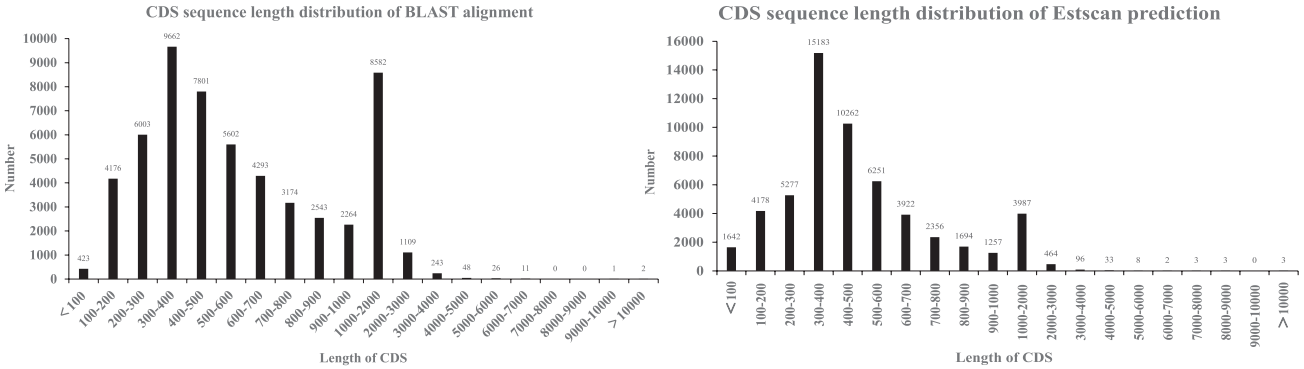

Fig. 6. CDS prediction of *Orychophragmus violaceus* (Left: BLAST alignment; Right: Estscan prediction).

Acknowledgments

This research was finally supported by National Natural Science Foundation of China (32001101), the project of Science and Technology of Guizhou Province (Qian Ke He Zhi Cheng [2019]2356), and the Doctor Foundation Project of Guizhou Normal University (2017).

Conflict of Interest

The authors declare no conflict of interest.

References

1. LIU C., HUANG Y., WU F., LIU W., NING Y., HUANG Z., TANG S., LIANG Y. Plant adaptability in karst regions. Journal of Plant Research, **134** (5), 889, **2021**.
2. CAO W., XIONG Y., ZHAO D., TAN H., QU J. Bryophytes and the symbiotic microorganisms, the pioneers of vegetation restoration in karst rocky desertification areas in southwestern China. Applied Microbiology and Biotechnology, **104** (3), 873, **2020**.
3. XU M., LI A., TENG Y., SUN Z., XU M. Exploring the adaptive mechanism of passiflora edulis in karst areas via an integrative analysis of nutrient elements and transcriptional profiles. BMC Plant Biology, **19** (1), 185, **2019**.
4. LI H., TANG X., YANG X., ZHANG H. Comprehensive transcriptome and metabolome profiling reveal metabolic mechanisms of *Nitraria sibirica* pall. to salt stress. Scientific Reports, **11** (1), 12878, **2021**.
5. DAI P., SUN G., JIA Y., PAN Z., TIAN Y., PENG Z., LI H., HE S., DU X. Extensive haplotypes are associated with population differentiation and environmental adaptability in Upland cotton (*Gossypium hirsutum*). Theoretical and Applied Genetics, **133** (12), 3273, **2020**.
6. ZHOU S., XING Z., LIU H., HU X., GAO Q., XU J., JIAO S., JIA K., JIN Y., ZHAO W., PORTH I., EL-KASSABY Y.A., MAO J. In-depth transcriptome characterization uncovers distinct gene family expansions for *Cupressus gigantea* important to this long-lived species' adaptability to environmental cues. BMC Genomics, **20** (1), 213, **2019**.
7. PANG Y., LI L., BIAN F. Photosynthetic and ultrastructural adaptability of *Anemone shikokiana* leaves to heterogeneous habitats. Brazilian Journal of Botany, **43** (4), 979, **2020**.
8. ASEEVA T.A., ZENKINA K.V. Adaptability of Spring Triticale Varieties in the Agroecological Conditions of the Middle Amur. Russian Agricultural Sciences, **45** (2), 112, **2019**.
9. YANG Q., JIANG Z., YUAN D., MA Z., XIE Y. Temporal and spatial changes of karst rocky desertification in ecological reconstruction region of Southwest China. Environmental Earth Sciences, **72** (11), 4483, **2014**.
10. YUAN S., GUO C., MA L.N., WANG R. Environmental conditions and genetic differentiation: What drives the divergence of coexisting Leymus chinensis ecotypes in a large-scale longitudinal gradient? Journal of Plant Ecology, **9** (5), rtv084, **2016**.
11. MOORE A.J., MOORE W.L., BALDWIN B.G., DANIEL O.B. Genetic and ecotypic differentiation in a californian plant polyploid complex (*Grindelia*, Asteraceae). Plos One, **9** (4), e95656, **2014**.
12. REN Y., LU Y., FU B. Quantifying the impacts of grassland restoration on biodiversity and ecosystem services in China: A meta-analysis. Ecological Engineering the Journal of Ecotechnology, **95**, 542, **2016**.
13. REN Y., LU Y., FU B., ZHANG K. Biodiversity and Ecosystem Functional Enhancement by Forest Restoration: A Meta-analysis in China. Land Degradation & Development, **28** (7), 2062, **2017**.
14. LUO P., HUANG B.Q., YIN J.M., CHEN Z.L., CHEN Y.H., LAN Z.Q. A new forage genetic resource *Orychophragmus violaceus* (L.) O.E. Schulz. Genetic Resources and Crop Evolution, **45** (6), 491, **1998**.
15. XING D., CHEN L., WU Y., ZWIAZEK J.J. Leaf physiological impedance and elasticity modulus in *Orychophragmus violaceus* seedlings subjected to repeated osmotic stress. Scientia Horticulturae, **276** (1), 109763, **2021**.
16. HANG H., WU Y. Effect of Bicarbonate Stress on Carbonic Anhydrase Gene Expressions from *Orychophragmus violaceus* and *Brassica juncea* seedlings. Polish Journal of Environmental Studies, **28** (3), 1135, **2019**.
17. JAVED Q., WU Y., XING D., ULLAH I., AZEEM A., RASOOL G. Salt-induced effects on growth and photosynthetic traits of *Orychophragmus violaceus* and its restoration through re-watering. Brazilian Journal of Botany, **41** (1), 29, **2017**.
18. LIU J., CAO W., RONG X., JIN Q., LIANG J. Nutritional characteristics of *Orychophragmus violaceus* in north china. Soil and Fertilizer Sciences in China, (1), 78, **2012**.
19. WANG R., WU Y., HANG H., LIU Y., XIE T., ZHANG K., LI H. *Orychophragmus violaceus* L., a marginal land-based plant for biodiesel feedstock: Heterogeneous catalysis, fuel properties, and potential. Energy Conversion and Management, **84** (June), 497, **2014**.
20. LUO P., LAN Z.Q., LI Z.Y. *Orychophragmus violaceus*, a potential edible-oil crop. Plant Breeding, **113** (1), 83, **2010**.
21. MA M., MEI Y. Research Status and Development Prospect of *Orychophragmus violaceus*. Journal of Anhui Agricultural ences, **40** (09), 5109, **2012**.
22. LIU C.Q., ZHU N.L., YANG S.X., DAI Y.H., CAO L. Protective effect of *Orychophragmuspine* I against oxidative damage in HepG2 cells induced by hydrogen peroxide. Modern Food Science and Technology, **33** (6), 19, **2017**.
23. DU S.Z., DAI Q., FENG B., JIN, W. The epsps gene flow from glyphosate-resistant *Brassica napus* to untransgene *B. napus* and wild relative species *Orychophragmus violaceus*. Acta Physiologiae Plantarum, **31** (1), 119, **2009**.
24. LI N.X., HUANG Y.L., LIANG G., JIANG M. Isolation and sequence analysis of a chalcone synthase gene *OvCHS* from *Orychophragmus violaceus*. Journal of Taizhou University, **32** (03), 42, **2010**.
25. VELEY K.M., BERRY J.C., FENTRESS S.J., SCHACHTMAN D.P., BAXTER I., BART R. High-throughput profiling and analysis of plant responses over time to abiotic stress. Plant direct, **1** (4), e00023, **2017**.
26. LUZ A., PRETTI I.R, BATITUCCI M. Comparison of RNA extraction methods for *Passiflora edulis* SIMS leaves. Revista Brasileira de Fruticultura, **38** (1), 226, **2016**.
27. HU Y., WANG K., HE X., CHIANG D.Y., PRINS J.F., LIU J. A probabilistic framework for aligning paired-end RNA-seq data. Bioinformatics, **26** (16), 1950, **2010**.
28. CONESA A., GOTZ S., GARCIA-GOMEZ J.M., TEROL J., TALON M., ROBLES M. Blast2GO: a universal tool

for annotation, visualization and analysis in functional genomics research. Bioinformatics, **21** (18), 3674, **2005**.
29. DERBYSHIRE M.K., GONZALES N.R., LU S., HE J., MARCHLER G.H., WANG Z., ARON M.B. Improving the consistency of domain annotation within the conserved domain database. Database, **2015**, bav012, **2015**.
30. ISELI C., JONGENEEL C.V., BUCHER P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proceedings. International Conference on Intelligent Systems for Molecular Biology, **99**, 138, **1999**.
31. REDDY R., HARISHBHAI M.R., HARENDRABHAI S.P., JAYANTI M., ATHAMARAM G.N., MANIVEL P., JITENDRA K., SUBUDHI P.K. Next generation sequencing and transcriptome analysis predicts biosynthetic pathway of sennosides from senna (*Cassia angustifolia* Vahl.), a non-model plant with potent laxative properties. Plos One, **10** (6), e0129422, **2015**.
32. MITSUNORI K., HIDETOSHI M., RUI Y., SEIYA I., SATORU M. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. Biostatistics, **17** (2), 1, **2016**.
33. MINORU K., SUSUMU G., YOKO S., MASAYUKI K., MIHO F., MAO T. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research, **42** (D1), D199, **2014**.
34. PANG M., SHAN Y., WANG F., YANG S. Protective effect of water extracts of *Orychophragmus violaceus* seeds on taa-induced acute liver injury in mice. China journal of Chinese materia medica, **45** (6), 1399, **2020**.
35. ZHANG L., DAI S. Genetic variation within and among populations of *Orychophragmus violaceus* (Cruciferae) in China as detected by ISSR analysis. Genetic Resources & Crop Evolution, **57** (1), 55, **2010**.
36. ZHANG K., WU Y., HANG H. Differential contributions of NO_3^-/NH_4^+ to nitrogen use in response to a variable inorganic nitrogen supply in plantlets of two brassicaceae species in vitro. Plant Methods, **15** (1), 86, **2019**.
37. WANG R., WU Y., HANG H., LIU Y., XIE T., ZHANG K., LI H. *Orychophragmus violaceus* L. a marginal land-based plant for biodiesel feedstock: heterogeneous catalysis, fuel properties, and potential. Energy Conversion and Management, **84**, 497, **2014**.