Plant Omics Journal POJ

POJ 8(6):493-499 (2015) ISSN:1836-3644

The first insight into transcriptome profile of herbaceous plant *Nervilia fordii* based on RNA-seq

Qionglin HUANG[1,2], Lingling LIANG[1], Rui HE[1]*, Xinye MA[1], Ruoting ZHAN[1], Weiwen CHEN[1]*

[1]Research Center of Chinese Medicinal Resource Science and Engineering, Guangzhou University of Chinese Medicine Key Laboratory of Chinese Medicinal Resource from Lingnan, Ministry of Education, Guangzhou 510006, P.R. China
[2]Guangdong Medical University, Zhanjiang 524023, P.R. China

Abstract

Nervilia fordii is a famous and valuable herbaceous plant that has been used to cure pulmonary and respiratory diseases for hundred years. However, the wild resource of *N. fordii* has been virtually exhausted, and the available approaches such as artificial cultivation and tissue proliferation cannot produce enough plants to satisfy the clinic and market. Utilization of biological technique to regulate some physiological processes of the species is a considerable method to solve the scarcity of *N. fordii*. To date, little is known about functional genes from *N. fordii*, especially the genes involved in biosynthesis pathways of effective components. Herein, a transcriptomic study was performed using leaf and corm of *N. fordii* as material. A total of 102,258,558 high-quality reads were produced by Illumina RNA-seq platform and finally 142,220 unigenes were assembled with an average length of 518 bp. The distinct genes were searched against an Nr database and 37.9% (53,970) of them had at least one hit. We also found 38,640, 110,029, and 28,970 of the unigenes assigned in COG, GO and KEGG databases, respectively. We discovered almost every important gene participated in flavonoids and terpenoids biosynthesis pathways and illustrated their expression difference between leaf and corm of *N. fordii*. We presented the first transcriptome insight for *N. fordii* and provide a firm foundation for genetic manipulation of the endangered plant.

Keywords: *Nervilia fordii* (Hance) Schltr.; RNA-seq; Transcriptome; Flavonoids and Terpenoids.
Abbreviation: COG_Cluster of Orthology; GO_Gene Ontology; HQ_Hhigh quality; KEGG_Kyoto encyclopedia of genes and genomes; NGS_ Next generation sequencing; Nr_Non-redundant.

Introduction

Nervilia fordii, which belongs to the Orchidaceae family, is a famous and valuable herbal medicine and has long been used as traditional Chinese medicine *Qingtiankui* in the *Lingnan* area (a district including Guangdong, Guangxi, Hainan and Fujian Provinces of China), for its significant effect in curing pediatric respiratory diseases. Due to excessive human harvest, wildlife resources of *N. fordii* have been endangered, and its rigorous requirements of germination and growth aggravate the situation (Chen, 2010; Chen and Xu, 2007). The species has been listed in "Catalogue of Rare and Endangered Plants Grown on Limestone in South China" and "International Trade Convention for Endangered Plants and Animals" (Wen et al., 1993). Until now, tissue proliferation and domestic cultivation have been attempted to ease the plight (Du et al., 2013; Du et al., 2005a; Du et al., 2005b). However, these methods encountered their bottleneck problems with seedling scarcity, seedling hardening, and cultivation expansion; the resource of *N. fordii* acquired from these approaches is far from satisfactory of the market and clinic demand.

Research on chemical constituents of *N. fordii* have also been carried out and revealed that flavonoid were among the effective compounds. For example, five flavonoid glycosides named nervilifordins F-J exhibited inhibitory effect on nitric oxide production in lipopolysaccharide activated RAW264.7 macrophages (Qiu et al., 2013; Zhou et al., 2009). The flavonoid derivative, Rhamnocitrin, directly inhibited six tumor strains including L1210, P338D1, Hela, B16, NG108-15 and Hele 7406 in a dose-dependent manner (Zhen et al., 2008). In addition, terpenoids was another major chemical ingredient reported in *N. fordii* and approximately 10 compounds were isolated (Huang et al., 2014; Wei et al., 2012; Zhao, 2006). Under the circumstances that the existing approaches were unable to solve the shortage, characterization of the genes participated in effective compounds biosynthesis in order to artificially increase the amount of these compounds at molecular level will contribute to overcoming the shortage and endangerment of *N. fordii*. However, little is known about function genes from *N. fordii*.

The identification of functional genes involved in flavonoids and terpenoids biosynthesis pathways has been achieved using various techniques, including RNA-seq based on NGS (Tang et al., 2011; Huang et al., 2012; Zheng et al., 2014). RNA-seq has shown great potential for high throughput sequencing and provides massive sequence in transcriptome with enormous depth and coverage to easily discover novel genes, splice junction, fusion transcripts and to make a comparison between samples or tissues (Garg and Jain, 2013; Van Verk et al., 2013; Wang et al., 2009). It has also been identified as an effective tool to discover functional genes for non-model plants (Guo et al., 2013; Hao et al., 2011; Zheng et al., 2014). Here, we carried out a transcriptomic study by means of Illumina RNA-seq platform and mined candidate genes involved in flavonoids and terpenoids biosynthesis of *N. fordii*.

Results

Illumina sequencing and de novo assembly

By Illumina RNA-seq platform, 47,668,726 and 524,589,832 clean reads were acquired from leaf and corm tissues, respectively, with total nucleotides 9,203,270,220 (9.20 Gb). And 26,341,686 and 33,2851,37 clean reads in leaf and corm were used for *de novo* assembly, respectively. The average read length, Q20 percentage (sequencing error rate < 1%), N (ambiguous bases) percentage and GC percentage listed in Table 1, revealed high sequencing quality in two tissues of *N. fordii*.

De novo of these HQ reads generated 95,761 contigs with mean size of 511 bp in leaf and 107,309 contigs were clustered with mean size of 531 bp in corm. From these contigs, 92,784 and 103,745 unigenes were assembled in leaf and corm respectively, and finally 142,220 unigenes with an average length of 518 bp and an N50 of 650 bp were acquired from this high-throughputs sequencing of *N. fordii*. Among 142,220 unigenes, 103,495 (72.77%) were shorter than 500 bp, 22,616 (15.90%) ranged from 500 to 1000 bp, 8,271 (5.82%) ranged from 1000 to 1500 bp, 4,107 (2.89%) distributed from 1500 to 2000 bp, and 3,371 (2.62%) were longer than 2000 bp. The distributions of these contigs and unigenes were shown in Fig. S1.

Gene function annotation

Functional annotation of unigenes was carried out by searching unigene sequences against public databases, Nr, Swissprot, GO, COG and KEGG (E-value $<10^{-5}$), and predicting the protein function from the annotation of the most similar proteins. These distinct genes were firstly searched in the Nr database using BLASTX, and 53,970 unigenes (37.9% of all unigenes) produced 836,804 hits that exceeded the E-value cutoff. Similarly, up to 40,712 unigenes were mapped to Swissprot database. GO is an international standardized database on classification of gene function, and provides a dynamically updated, controlled vocabulary to comprehensively describe the properties of genes and their products in any organism. Three ontologies-biological process, cellular components and molecular function-constitute this database, and basic units composed of the ontology are named as "GO term". GO annotation results of *N. fordii* unigenes were acquired according to Nr annotation, and then conducted GO functional classification for unigenes and examined the macro-level distribution of gene functions using WGEO software. In total, 110,029 unigenes have their GO-annotations and were assigned 8,953 GO-term hits by BLAST2GO, and the terms were classified into the three main GO categories and 44 sub-categories (Fig. 1). In each of three main categories, "metabolic process" (11, 109 unigenes), "cell" (16,140 unigenes) and "binding" (10,770 unigenes), respectively, were the dominant. Moreover, "cellular process", "cell part", and "catalytic activity" also occupied a high-percentage of unigenes and only a few distinct genes were summarized into "nitrogen utilization", "cell junction", and "metallochaperone activity".

COG is a database that includes classifications of orthologous gene products. Each protein in COG is assumed to be evolved from an ancestor protein, and the database contains the coding proteins of complete genomes and the evolutionary relationship of bacteria, algae and eukaryotes. In order to predict and classify the possible function of *N. fordii* unigenes, we did a search against COG database. In total, 38,640 unigenes were annotated in COG, and categorized to 25 classifications (Fig.2). Among 25 functional categories, "general function prediction only" occupied the most, 5,789 unigenes (about 15.0% of all annotated unigenes in COG), followed by "translation, ribosomal structure and biogenesis" (3,554; 9.20%), "transduction" (3,211; 8.31%). And the categories: "Nuclear structure" (4; 0.01%), "Extracellular structures" (15; 0.04%) and "RNA processing and modification" (247; 0.64%), represented the smallest groups in this annotation. Importantly, 998 unigenes were assigned into the group "secondary metabolites biosynthesis, transport and catabolism"(Q).

To further understand the biological function of the distinct genes, pathway annotation against KEGG database was carried out. The KEGG pathway database contains information on networks of intracellular molecular interactions and their organism-specific variation. By mapping the annotated unigenes to the available pathways in KEGG, a total number of 28,970 unigenes were assigned into 281 pathways (Table S1). The highly represented pathway was "metabolic pathway" (7,639 members), followed by "biosynthesis of secondary metabolites" (3,452 members) and "ribosome" (1,894 members). We also found 2,111 unique genes corresponded to pathways related to the biosynthesis of secondary metabolites (Table 2). Among them, the cluster for "Phenylpropanoid biosynthesis [path: ko00940]" consists of the largest group (479, 22.7%), followed by "zeatin biosynthesis [path: ko00908]" (375, 17.7%) and "Limonene and pinene degradation [path: Ko00903]" (255, 12.1%).

A total of 110,029 unigenes were identified from one or more of the databases mentioned above, indicating that they're high conservation on function. These annotations also provide a valuable resource for discovering functional genes and investigating specific process and pathways of *N. fordii*.

Candidate genes involved in flavonoids biosynthesis

Flavonoids, including anthocyanidins, flavonols, flavones *etc*, are a group of plant polyphenolic secondary metabolites showing a common three ring chemical structure (C_6-C_3-C_6). The basic skeleton of all flavonoids starts from three molecules of manlonyl-CoA and one of 4-counmaroyl-CoA. Chalcone synthase (CHS) and Chalcone isomerase (CHI) are the enzymes involved in the two-step condensation, and create a flavanone named naringenin. Then, the oxidation and catalysis of the latter compound by special enzymes begins to synthesize different flavonoids and anthocyanidins are the end product of the biosynthesis pathway. In the transcriptome of *N. fordii*, Unigenes encoded 12 enzymes involved in flavonoids biosynthesis were identified and a putative biosynthesis pathways of flavonoids in *Nervilia fordii* were presented based on the annotated genes (Fig. 3 and Table S2).

Candidate genes participated in terpenoids biosynthesis

Terpenoids are derived from C5 isoprene units through a "head-to-tail" connection. The conjunction of a different number of C5 isoprene units brings about various intermediates, such as IPP (C5 unit), DMAPP (C5 unit), GPP (C10 unit), and FPP (15 unit), which form the carbon skeletons of the different terpenoids. IPP, along with its isomer DMAPP, are important intermediates in terpenoids backbone formation; and both intermediates can be synthesized through the MVA pathway in the cytoplasm, or the MEP pathway in the plastid. Thirteen unigenes participating in the upstream process of terpenoids biosynthesis were identified in *N. fordii* transcriptome, including 6 in MVA pathway and 7 in MEP pathway (Fig. 4 and Table S3). Both MVA and MEP pathways produce the C5 unit IPP, which can be transformed into its isomer, DMAPP, by IDI (Isopentenyl diphosphate isomerase). In the downstream

Table 1. Summary for transcriptome signatures from two tissues *N. fordii*.

	Leaf	Corm
Total number of clean reads	47,668,726	54,589,832
Total nucleotides/nt	4,290,185,340	4,913,084,880
Average read length/bp	90	90
Q20 percentage/%	96.12	96.05
N percentage/%	0.00	0.00
GC percentage/%	43.93	44.26
Number of clean reads used for de novo assembly	26,341,686	33,2851,37
Assembly percentage of clean reads/%	55.26	60.97

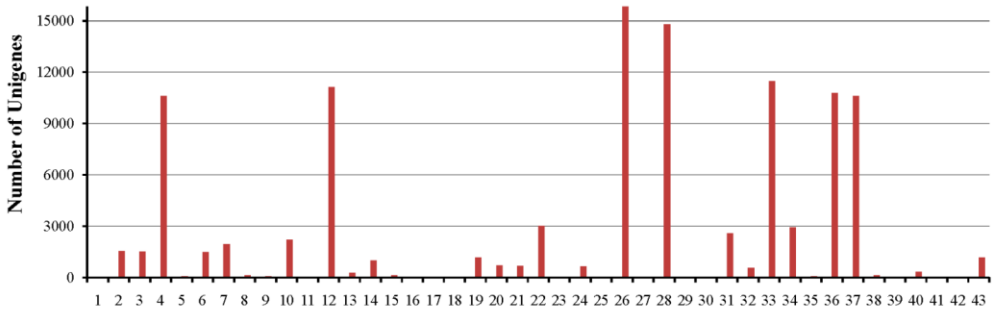

Fig 1. Chart presentations of Gene Ontology classification. 100,029 unigenes were assigned into three main ontologies of GO database, biological process, cellular components and molecular function. The sub-ontologies from 1 to 25 belong to biological process, the ones from 26 to 34 belong to cellular component, and the ones from 35 to 43 belong to molecular function. 1. Biological adhesion; 2. Biological regulation; 3. Cellular component organization or biogenesis; 4. Cellular process; 5. Death; 6. Developmental process; 7. Establishment of location; 8. Growth; 9. Immune system process; 10. Localization; 11. Locomotion; 12. Metabolic process; 13. Multi-organism process; 14. Multicellular organismal process; 15. Negative regulation of biological process; 16. Nitrogen utilization; 17. Pigmentation; 18. Positive regulation of biological process; 19. Regulation of biological process; 20. Reproduction; 21. Reproductive process; 22. Response to stimulus; 23. Rhythmic process; 24. Signaling; 25. Viral reproduction; 26. Cell; 27. Cell junction; 28. Cell part; 29. Extracellular region; 30. Extracellular region part; 31. Macromolecular complex; 32. Membrane-enclosed lumen; 33. Organelle; 34. Organelle part; 35. Antioxidant activity; 36. Binding; 37. Catalytic activity; 38. Enzyme regulator activity; 39. Metallochaperone activity; 40. Molecular transducer activity; 41. Protein binding transcription factor activity; 42. Receptor activity; 43. Transporter activity.

Table 2. The unigenes involved in secondary metabolites.

Pathway of secondary metabolites	Number of unigenes
Anthocyanin biosynthesis	6
Betalain biosynthesis	3
Brassinosteroid biosynthesis	25
Caffeine metabolism	13
Carotenoid biosynthesis	137
Diterpenoid biosynthesis	64
Flavone and flavonol biosynthesis	72
Flavonoid biosynthesis	218
Glucosinolate biosynthesis	12
Indole alkaloid biosynthesis	26
Limonene and pinene degradation	255
Novobiocin biosynthesis	25
Phenylpropanoid biosynthesis	479
Stilbenoid, diarylheptanoid and ginerol biosynthesis	210
Streptomycin biosynthesis	43
Terpenoid backbone biosynthesis	63
Tetracycline biosynthesis	21
Tropane, piperidine and pyridine alkaloid biosynthesis	62
Zeatin biosynthesis	375

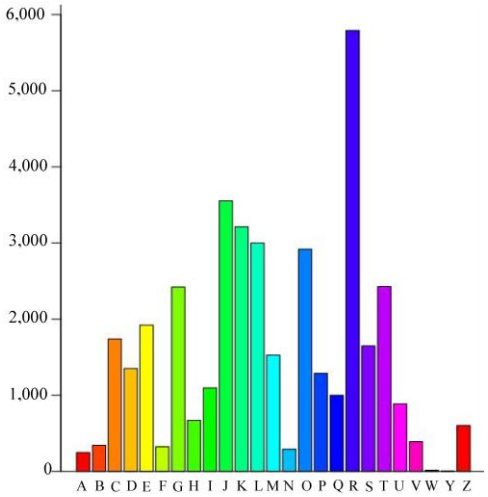

Fig 2. Histogram presentations of clusters of orthologous groups. A. RNA processing and modification. B. Chromatin structure and dynamics. C. Energy production and conversion. D. Cell cycle control, cell division, chromosome partitioning. E. Amino acid transport and metabolism. F. Nucleotide transport 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.

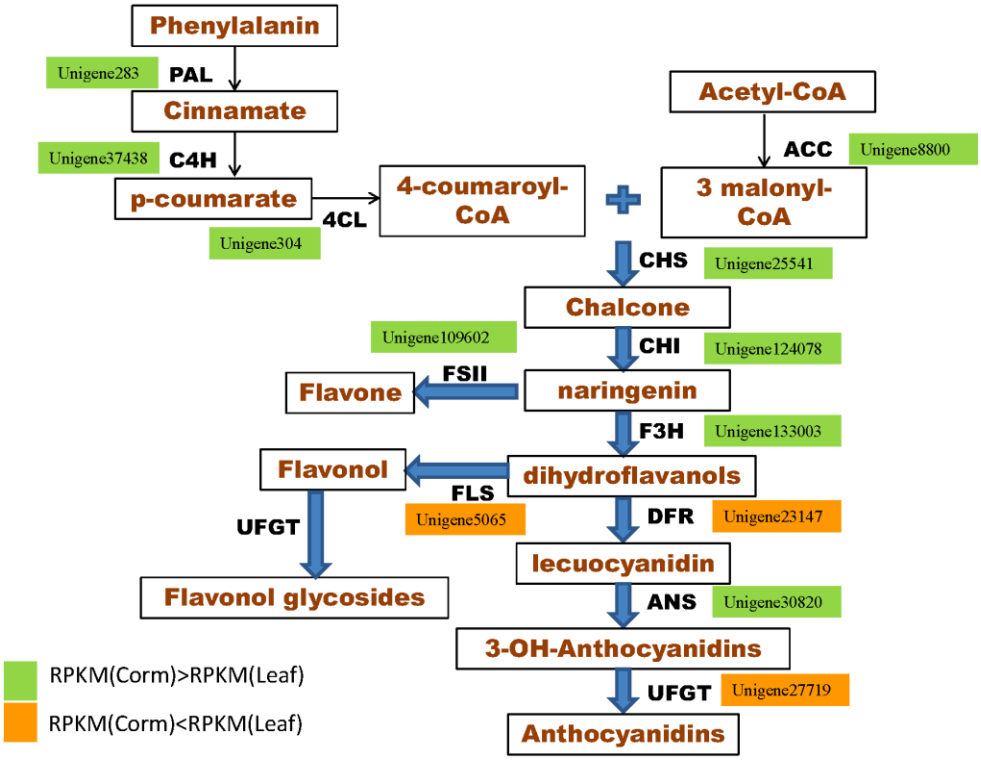

Fig 3. Unigenes involved in flavonoids biosynthesis annotated in transcriptome of *N. fordii*. The unigenes in green boxes represented that their expression levels in corm were higher than in leaf, and the unigenes in yellow boxes meant that their expression level in corm were lower than in leaf. PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumaroyl:CoA ligase; ACC, acetyl-CoA catboxylase; CHS, chalcone synthase; CHI, chalcone isomerase; FSII, flavones synthase II; F3H, flavanone 3-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin sythase; UFGT, UDP glucose flavonoid 3-O glucosyltransferas.

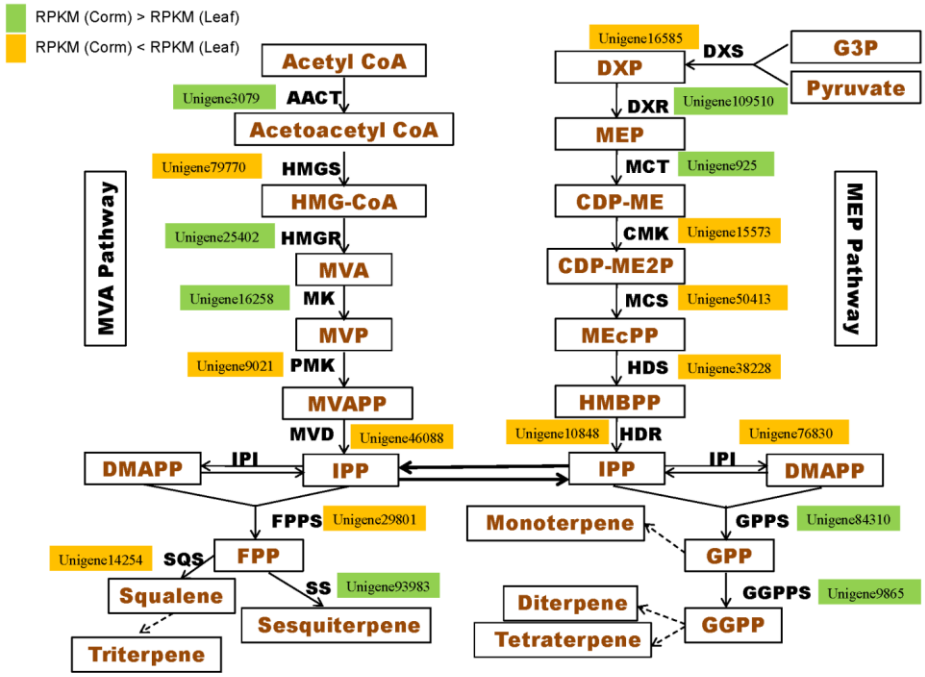

Fig 4. Unigenes involved in terpenoids biosynthesis annotated in transcriptome of *N. fordii*. The unigenes in green boxes represented that their expression levels in corm were higher than in leaf, and the unigenes in yellow boxes meant that their expression level in corm were lower than in leaf.The step with dash arrow represented that the enzymecatalyzed this step was not annotated in this study.AACT, Acetyl-CoA acetyltransferase; HMGR, 3-hydroxy-3-methylglutaryl- coenzyme A reductase; HMGS, 3-hydroxy-3-methylglutaryl-coenzyme A synthase; MK, mevalonate kinase; PMK, phosphomevalonate kinase; MVD, diphosphomevalonate decarboxylase; DXS, 1-deoxy-D-xylulose-5-phosphate synthase; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; MCT, 3-C-methyl-D-erythritol 3-phosphate cytidylyltransferase; CMK, 3-(cytidine 5'-diphospho)-3-C-methyl-D-erythritol kinase; MCS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, -hydroxy-3-methylbut-2-enyl diphosphate synthase; HDR, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase; IPI, isopentenyl diphosphate isomerase; FPPS, farnesyl pyrophosphare synthase; GPPS, geranyl diphosphat synthase; GGPPS, geranyl geranyl diphosphat synthase; SS, sesquiterpenoid synthase; SQS, squalene synthase.

pathway, IPP and DMAPP are assembled into GPP, FPP and GGPP by a series of prenyl transferases, including GPPS, FPPS and GGPPS, and finally formed various terpenoids compounds. In this study, 5 unigenes were mapped into the FPPS, GPPS, GGPPS, SQS and SS respectively (shown in Fig. 4 and Table S3). However, only two synthases of final terpenoid products-sesquiterpenoid synthase (SS) and squalene synthase (SQS)-were annotated, speculating that sesquiterpenoid, sterides and saponin were the major compositions of terpenoids in *N. fordii*.

Discussion

RNA-seq, a high-throughout mRNA sequencing technology, has been considered as fast, efficient and cost-effective approach to characterize the poly $(A)^+$ transcriptome. It is especially appropriate for gene discovery and expression profiling in non-model organisms that lack genomic sequences. To date, more and more RNA-seq studies on traditional Chinese medicines, which mostly lack genomic information, were carried out to demonstrate transcriptome profile and mine functional gene (Hao et al., 2012; Yuan et al., 2012; Zheng et al., 2014; Qi et al., 2015). In this study, we applied RNA-seq technology for *N. fordii* transcriptome profiling, in which the poly $(A)^+$ transcriptome was sequenced on the Illumina platform. We obtained 9.2 G bp coverage with 10.2 million HQ reads from leaf and corm tissue of *N. fordii*. We generated a total of 142,220 unigenes (>100 bp) by de novo assembly. Among them, 110,029 assembled unigenes were annotated.

These distinct genes were assigned not only gene or protein name description, but also predicted coding sequences, gene notology terms and metabolic pathways. Detailed functional information is crucial in order to deepen our comprehension on overall expression profiles of *N. fordii*. This was the first study to explore the mRNA sequences and annotation of *N. fordii* in large scale, which provided abundant data for functional genes researches of *N. fordii*. With the help of RNA-seq transcriptome, interest in functional genes involved in metabolic biosynthesis and establishing modern metabolic and biosynthesized system for effective compounds of herbal medicines increased, which aimed to breed high-quality, high-yield, and anti-stress medicinal plants via utilization and manipulation of these genes. The medicinal quality of *N. fordii* in large part depends on its metabolic profiles of effective compounds, and we focused on flavonoids and terpenoids biosynthesis for additional analyses. Importantly, we were able to find almost all metabolic genes involved in these biosynthesis pathways. We are confident that our transcriptome dataset is a valuable supplement to the publicly available genomic information of *N. fordii*, even entire *Nervilia* genus. However, several of the unigenes involved in flavonoids and terpenoids biosynthesis pathway annotated in this study merely contained partial coding sequences. Next, we will obtain the full-length, open reading fragment of these genes and confirm their function in flavonoids and terpenoids synthesis via plant genetic engineering, a series of techniques including rapid amplification of cDNA ends and reverse transcription polymerase chain reaction. Once plant materials were available,

the expression differences of target genes in various tissues of *N. fordii* should be further validated using qRT-PCR. The identification and functional characterization of genes involved in flavonoids and terpenoids biosynthesis pathway not only facilitates foundational studies but also allow us to improve the effective flavonoids and terpenoids production of *N. fordii* via metabolic engineering, which will enhance the medicinal quality and further alleviate the resource shortage of *N. fordii*.

Materials and Methods

Plant materials

Mature *Nervilia fordii* was collected from Mashan county, Nanning city, Guangxi Zhuang Autonomous Region, China. The leaf and corm tissues were separated and washed under running water to remove soil and other attachments, and then cut into small pieces and immediately frozen in liquid nitrogen until further protocols.

RNA isolation

Total RNA were extracted from leaf and corm of the same plant using a previously reported CTAB method (Chen, 2011), and treated with RNase-free DNase for 30 minutes at 37 ^{o}C to remove residual DNA. The integrity, purity, and concentration of the total RNA were analyzed using ultraviolet spectroscopy.

cDNA library preparation and Illumina sequencing

Illumina sequencing was conducted at the Beijing Genomics Institute (BGI), Shenzhen, Guangdong province, China (http://www.genomics.cn/index.php) according to the manufacturer's instructions (Illumina, San Diego, CA, USA) (Grabherr et al., 2011).

De novo assembly and sequence clustering

The clean reads called raw data were acquired from original sequencing reads and saved in fastq format. The application of Trinity software (Haas et al., 2013) completed the de novo assembly of the transcriptome into unigenes. These assembled unigenes can be taken into further processing of sequence splicing and redundancy removing with sequence clustering software to acquire non-redundant unigenes. The utilization of Blastx (E-value < 10^{-5}) revealed alignment between unigenes and public protein databases including Nr, Swissprot, COG and KEGG, and the direction of sequences were determined by the best aligning results. Additionally, ESTScan software (Iseli et al., 1999) was used to predict the direction of the unigenes that could not be mapped into those in any of the public databases mentioned above.

Gene function annotation

To achieve the best function annotations, the unigenes were mapped to protein databases such as Nr, Swissprot, CGO and KEGG, and nucleotide database Nt (E-value < 10^{-5}). By means of Nr annotation, Blast2 GO program (Conesa et al., 2005) was applied to gain gene ontology annotation of unigenes and then WEGO software (Ye et al., 2006) was used to perform GO functional classification for unigenes. The application of KEGG mapping (Kanehisa and Goto, 2000) determined metabolic pathway annotation of unigenes.

To further confirm the function of these genes annotated into flavonoids and terpenoids biosynthesis pathways, the unigene sequence were translated into amino acid sequence and submitted into Genbank Conserved Domain Database (CDD) to predict their conserved domains. Multiple alignments of nucleotide sequences were also performed to understand the homology between the annotated unigenes from *N. fordii* transcriptome and the identified counterpart from other plants submitted into Genbank. Finally, the unigenes annotated into flavonoids and terpenoids biosynthesis was categorized for *N. fordii* based on general plant flavonoids and terpenoids biosynthesis patterns, which have been fully clarified (Mahmoud and Croteau, 2001; Nishihara and Nakatsuka, 2011; Ma et al., 2012; Falcone Ferreyra et al., 2012).

Unigenes expression analysis

The expression levels of unigenes were evaluated by the method of reads per kb per million reads (RPKM) (Mortazavi et al., 2008). The RPKM value of a unigene was calculated using the formula as follows: RPKM= (1,000,000·C) /(N·L·1,000), where C is the number of reads that uniquely mapped to an objective unigene, N is the total number of reads that uniquely mapped to all genes, and L is the number of bases in the objective unigene. This method eliminates the influence of different length and sequencing level on the calculation of gene expression (Zhang et al., 2014). Thus, the calculated gene expression can be directly used for comparing the difference of gene expression between leaf and corm tissues of *N. fordii*.

Conclusion

To the best of our knowledge, this is the first time that the Illumina RNA-seq platform was applied to investigating the transcriptome signatures of *N. fordii*. Under the circumstance that the genome sequence of *N. fordi*i is still unavailable, the transcriptome analysis has provided 142,220 unigenes, among which 77.2% are annotated in one or more of Nr, Swissprot, GO, COG and KEGG databases. We identified the transcripts corresponding to genes associated with flavonoids biosynthesis, terpenoids biosynthesis. The transcriptome profile will provide a foundation for further genomics research on this species, even entire *Nervilia* genus.

Acknowledgements

This work was supported by special fund project of college discipline and specialty construction in Guangdong province of China (2013CXZDA011), scientific research foundation for returned scholars, Ministry of Education of China (NO.[2009]1001) and the PhD start-up fund of natural science foundation of Guangdong Province in China (2015A030310519). We thank Mr. Yunfeng Huang from Guangxi institute of Chinese Medicine and Pharmaceutical Science, P.R. China for his kind help with sample collection and authentication.

References

Chen W (2010) Materia medica in Lingnan, vol 2. Guangdong Science and Technology Press, Guangzhou, China.

Chen W, Xu H (2007) Research on material medica in Lingnan. Guangdong Science and Technology Press, Guangzhou, China.

Chen X (2011) *De novo* characterization of hazelnut floral bud transcriptome using Solex sequencing and expression profiling analysis of cold-regulated genes. Chinese Academy of Forest.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674-3676.

Du Q, Chen F , Gong X (2013) Study on aseptic germination of Nervilia fordii Seeds. J Guangzhou Univ Trad Chin Med. 30: 233-235.

Du Q, Chen W , Wang Z (2005a) Study on tissue culture and plant regeneration of Nervilia fordii. China J Chin Materia Medica. 30(11):812-814.

Du Q, Xu H, Wang Z (2005b) Primary investigation on growth status of artificial cultivation Nervilia fordii. J Chin Med Materia. (10):869-870.

Falcone Ferreyra ML, Rius SP, Casati P (2012) Flavonoids: biosynthesis, biological functions, and biotechnological applications. Front Plant Sci. 3:222.

Garg R, Jain M (2013) RNA-Seq for transcriptome analysis in non-model plants. Methods Mol Biol. 1069:43-58.

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

Guo X, Li Y, Li C, Luo H, Wang L, Qian J, Luo X, Xiang L, Song J, Sun C, Xu H, Yao H, Chen S (2013) Analysis of the Dendrobium officinale transcriptome reveals putative alkaloid biosynthetic genes and genetic markers. Gene. 527:131-138.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8:1494-1512

Hao D, Ge G, Xiao P, Zhang Y, Yang L (2011) The first insight into the tissue specific taxus transcriptome via Illumina second generation sequencing. PloS One. 6:e21220.

Hao D, Ma P, Mu J, Chen S, Xiao P, Peng Y, Huo L, Xu L, Sun C (2012) De novo characterization of the root transcriptome of a traditional Chinese medicinal plant Polygonum cuspidatum. Sci China Life Sci. 455:452-466.

Huang GK, Qiu L, Jiao Y, Xie JZ, Zou LH (2014) A new labdane diterpenoid glycoside from Nervilia fordii. Acta Pharmaceutica Sinica. 49: 652-655.

Huang L, Yang X, Sun P, Tong W, Hu S (2012) The First Illumina-Based De Novo Transcriptome Sequencing and Analysis of Safflower Flowers. PLoS One. 7: e38653.

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. 1999:138-148.

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27-30.

Ma Y, Yuan L, Wu B, Li X, Chen S, Lu S (2012) Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. J Exp Bot. 7: 2809- 2823.

Mahmoud SS, Croteau RB (2001) Metabolic engineering of essential oil yield and composition in mint by altering expression of deoxyxylulose phosphate reductoisomerase and menthofuran synthase. Proc Natl Acad Sci USA. 98:8915-8920.

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

Nishihara M, Nakatsuka T (2011) Genetic engineering of flavonoid pigments to modify flower color in floricultural plants. Biotechnol Lett. 33:433-441.

Qi J, Sun P, Liao D, Sun T, Zhu J, Li X (2015) Transcriptomic analysis of American ginseng seeds during the dormancy release process by RNA-Seq. PLoS One. 10:e0118558.

Qiu L, Jiao Y, Xie JZ, Huang GK, Qiu SL, Miao JH, Yao XS (2013) Five new flavonoid glycosides from Nervilia fordii. J Asian Nat Prod Res. 15:589-599.

Tang Q, Ma X, Mo C, Wilson IW, Song C, Zhao H, Yang Y, Fu W, Qiu D (2011) An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics, 12: 343.

Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci. 18:175-179.

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

Wei LB, Chen JM, Ye WC, Yao XS , Zhou GX (2012) Three new cycloartane glycosides from Nervilia fordii. J Asian Nat Prod Res. 14:521-527.

Wen H, Xu Z, Villa-Lobos J, Skog LE (1993) A list of threatened limestone plants in south of China. Guihaia. 13:110-127.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J (2006) WEGO: a web tool for plotting GO annotations. Nucleic acids research 34 (Web Server issue):W293-297.

Yuan Y, Song L, Li M, Liu G, Chu Y, Ma L, Zhou Y, Wang X, Gao W, Qin S, Yu J, Wang X, Huang L (2012) Genetic variation and metabolic pathway intricacy govern the active compound content and quality of the Chinese medicinal plant Lonicera japonica thumb. BMC Genomics. 13:195.

Zhang F, Wang Z, Dong W, Sun C, Wang H, Song A, He L, Fang W, Chen F, Teng N (2014) Transcriptomic and proteomic analysis reveals mechanisms of embryo abortion during chrysanthemum cross breeding. Sci Rep. 4:6536.

Zhao J (2006) Identification of chemical constituents from *Nervilia fordii*. Int J Appl Chem. 2(3):201-208.

Zhen H, Zhou Y, Yuan Y, Liang C, Qiu Q, Zhong Z , Zhang W (2008) Study on the anticancer effect in vitor of flavonoids compounds obtained from Nervilia fordii (Hance) Schltr. Chin J Trad Med Formu. 14:36-38.

Zheng X, Xu H, Ma X, Zhan R, Chen W (2014) Triterpenoid saponin biosynthetic pathway profiling and candidate gene mining of the Ilex asprella root using RNA-Seq. Int J Mol Sci. 15:5970-5987.

Zhou GX, Lu CL, Wang HS, Yao XS (2009) An acetyl flavonol from *Nervilia fordii* (Hance) Schltr. J Asian Nat Prod Res. 11:498-502.