ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

Study on Active Components of Four Rhododendron Species based on Data Mining

Yun He*

Department of Operation Management, Fuwai Central Cardiovascular Hospital, Zhengzhou 450000, China

Abstract

Based on transcriptomic data and gas chromatography-mass spectrometry (GC-MS), the active components of Rhododendron Flower tissue were extracted. Based on the transcriptome sequencing data of four Rhododendron species published in NCBI database, the transcripts were assembled by splicing strategy without reference genome, and the gene expression profiles of active components in Rhododendron species were established by annotation of go, KEGG, panther and other databases. At the same time, GC-MS technology was used to detect the tissue extract of Rhododendron, to verify the composition of plant active components. Through differential expression analysis and annotation of transcriptome, 34332 unigenes of Rhododendron fortunei Lindl, 36280 unigenes of Rhododendron molle, 50076 unigenes of Rhododendron mariesii and 82356 unigenes of Rhododendron simsii were identified. In addition, more than 80 kinds of metabolites were detected by GC-MS.

Keywords

Rhododendron; Transcriptome; Active Ingredient; Data Mining.

1. Introduction

Azalea, also known as Yingshan red and pomegranate, is a common name of Rhododendron, which is a common name for Rhododendron, about 960 species. Rhododendron has a wide range of vertical distribution. It can be seen that the vast majority of Rhododendron plants are distributed between 500-3800 meters above sea level in various vegetation zones from low altitude to high altitude, which has high ornamental value, medicinal value and economic value [1].

There are many Rhododendron plants with high medicinal value, among which the ones included in Chinese materia medica are of great value. There are 49 kinds of medicinal plants. The main chemical components of non-toxic medicinal plants of Rhododendron are flavonoids, terpenoids, volatile oil, etc. Quercetin derivatives and Rhododendron are widely distributed, and have good expectorant and antitussive effects, which are mainly used in the treatment of chronic bronchitis [4]. The research of Rhododendron toxic medicinal plants such as Rhododendron molle is more in-depth [6]. The toxicity of Rhododendron molle was recorded in compendium of Materia Medica in Ming Dynasty: "flowers, roots and leaves are highly toxic. Sheep eat their leaves and die by wandering. Some people used their roots to drink wine, so they died. The main toxic component of this kind of medicinal plants is resveratrol type diterpenoid toxin, which is widely used in the treatment of traumatic injury and rheumatoid arthritis pain. It has the effect of promoting blood circulation, dispersing blood stasis, detumescence and pain relief. Its usage is mostly for external use, and the dosage should be strictly controlled for internal use [8-9].

With the development of high-throughput sequencing technology, it is widely used in plant research. At present, high-throughput sequencing data of several rhododendrons are uploaded

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

in SRA of NCBI database [10]. In this study, we selected the transcriptome sequencing data of four Rhododendron species, namely Rhododendron fortunei Lindl, Rhododendron molle, Rhododendron mariesii and Rhododendron simsii. The transcripts were spliced into unigenes and the expressed proteins were predicted. The functional annotation was carried out by homologous sequence alignment and protein domain alignment, and the gene expression sequences related to active components were mined.

2. Materials and Methods

2.1. Data Sources

SRA database through NCBI (https://www.ncbi.nlm.nih.gov/sra/?term=) download the original transcriptome sequencing data of Rhododendron fortunei, Rhododendron molle, Rhododendron mariesii and Rhododendron simsii. The sequencing platform was Illumina HiSeq 2500, and the sequencing method was pair-end.

2.2. Transcriptome Sequencing Data Processing

2.2.1. Quality Control of Original Data

The software fastqc was used to evaluate the quality of transcriptome data of four Rhododendron species, including basic statistical information, GC content and adaptor content. According to the score of the software, it is decided whether to take the treatment of de jointing and low quality sequence. Follow up use fastx_toolkit according to the evaluation results for data quality control processing, get clean data.

2.2.2. Assembly and Processing of Transcriptome Sequences

Because there are no reference genomes in four Rhododendron species, the assembled sequence files can be obtained by using Trinity (v2.8.5) software. At the same time, Trinity software has many built-in data processing programs, which can calculate the number of reads on the assembled sequence, using the RSEM method.

2.2.3. Functional Annotation Identification of Transcriptome Sequences

The open reading frame (ORF) and possible coding region of the assembled transcriptome were predicted by TransDecoder (v5.5.0). After that, ORF can be identified by blast or Pfam searching the homologous sequence and domain information of known proteins. The databases used for alignment include Swiss-prot and Pfam. Finally, interproscan was used to annotate the transcriptome, mainly referring to Panther database. In addition, metascape analysis platform was used for enrichment pathway analysis.

2.3. Joint Analysis of Protein Data of Rhododendron

Based on the protein sequence information of Rhododendron fortunei, Rhododendron molle, Rhododendron mariesii and Rhododendron simsii in NCBI protein database, the known protein expression in each species was identified by blastx program, and the joint analysis was carried out according to the comparison results.

2.4. Expression and Identification of Active Components in Rhododendron

The active components in the flower tissue of Rhododendron molle were detected by gas chromatography-mass spectrometry (GC-MS). The experimental materials were dried flowers of Rhododendron molle; In this experiment, methanol, ethanol and benzyl alcohol were used for evaporation extraction, and then the concentrated extract was injected into the machine.

GC program: using elastic quartz capillary column hp-5ms (60m × two hundred and fifty μm × zero point two five μm) The carrier gas was high purity helium, the flow rate was 1 ml / min, the separation ratio was 10:1, and the temperature program of gas chromatography was from

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

50 ° C to 100 ° C, starting with 8 ° C / min to 250 ° C. Keep for 2 minutes, then use 5 minutes ° C / min to 280 ° C. Keep it for 5 minutes.

MS program: the scanning mass range is 30amu-600amu, the ionization voltage is 70ev, and the ionization current is 150 μA electron ionization (EI). The ion source and quadrupole temperature are set to 230 ° C and 150 ° C.

Then, the active components in Rhododendron plants were summarized by GC-MS detection results and related literature retrieval. The expression of the relevant active components and the encoded sequence information were identified according to the transcriptome data of four Rhododendron plants.

3. Results

3.1. Transcriptome Splicing of Four Rhododendron Species

The transcriptome data of four Rhododendron species were processed by using no reference genome splicing method. The data are from SRA database of NCBI, which are SRR5247112, SRR5247113, SRR5247114 and SRR7686809. The transcriptome of Rhododendron fortunei, Rhododendron molle, Rhododendron mariesii and Rhododendron simsii were 56.9M, 58.4M, 80.5M and 91.1M respectively. Table 1 is the basic statistics of splicing transcripts, including the number of transcripts and Unigene genes, N50 and GC content. It can be seen that GC content is in line with the quality control standard. Figure 1-Figure 4 shows the length distribution of four species of Rhododendron.

Table 1. Statistics of transcriptome splicing in four Rhododendron species

Rhododendron	Rhododendron fortunei Lindl.	Rhododendron molle (Blume) G. Don	Rhododendron mariesii Hemsl.&E.H.Wilson	Rhododendron simsii Planch.
transcripts	48996	52652	74238	111501
unigenes	34332	36280	50076	82356
Contig N50	1777	1700	1757	1379
GC	44.44	44.66	43.84	45.87

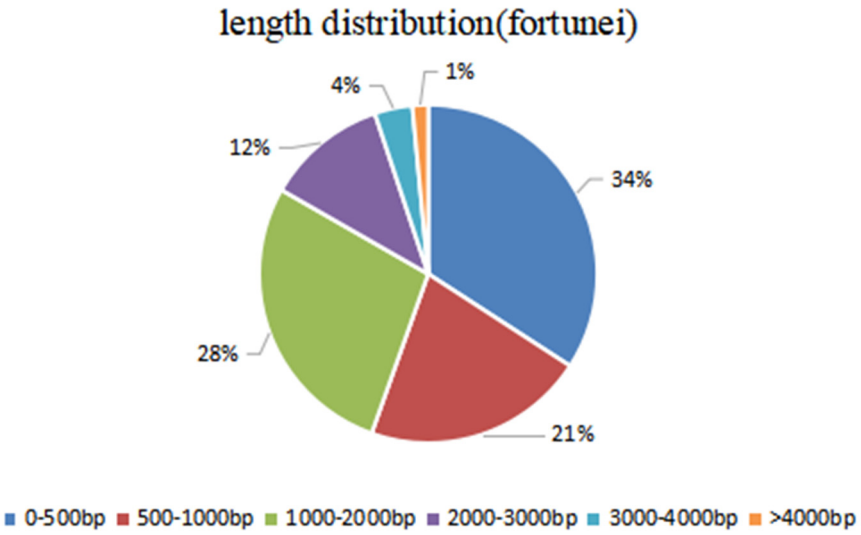

Figure 1. Length distribution of tissue transcripts of Rhododendron fortune

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

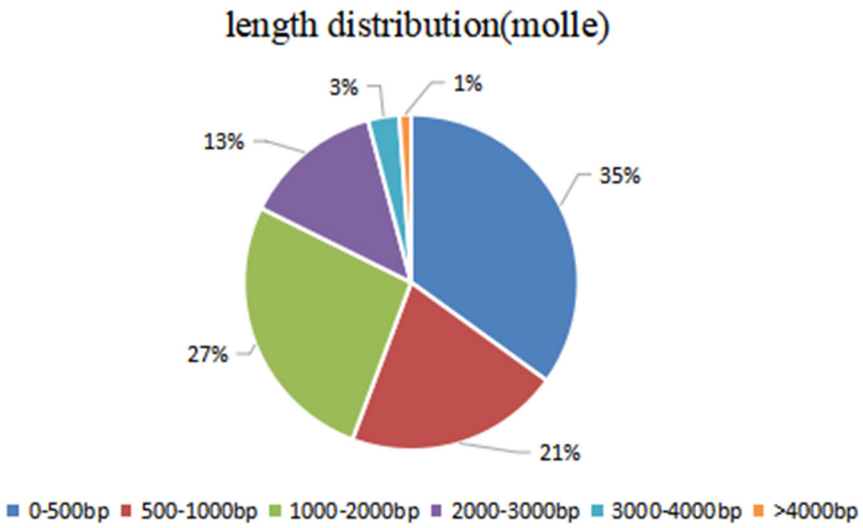

Figure 2. Length distribution of transcripts in Rhododendron mole

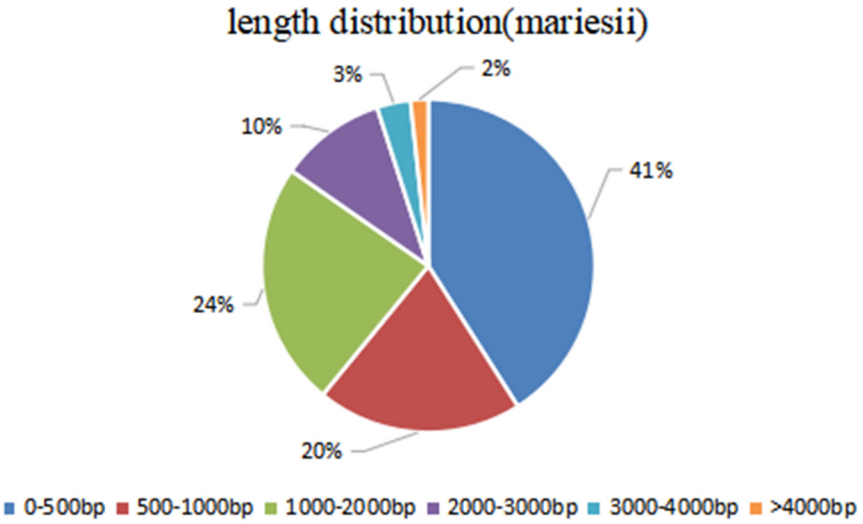

Figure 3. Length distribution of tissue transcripts of Rhododendron mariesii

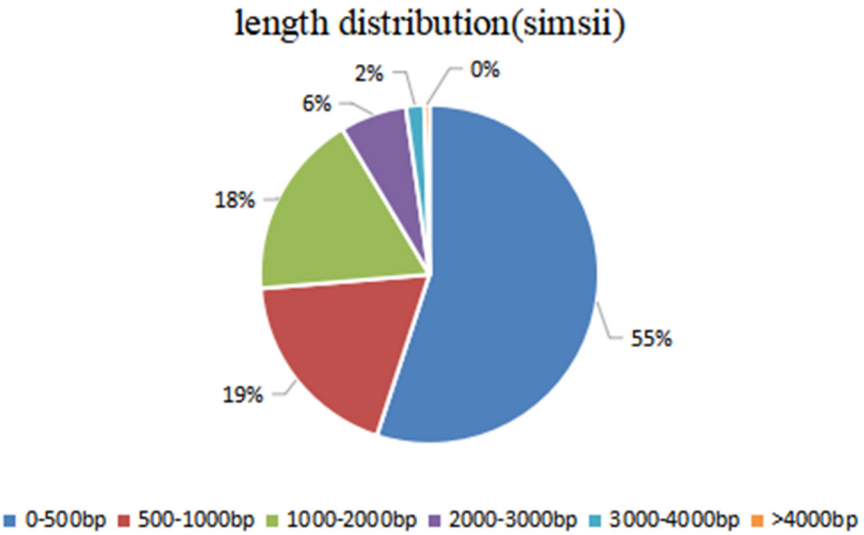

Figure 4. Length distribution of transcripts in Rhododendron simsii

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

Subsequently, the transcriptome expression of four Rhododendron species was analyzed by RSEM. Finally, the expression of different transcripts and their subtypes were counted. According to the number of transcripts expressed in each Rhododendron, we selected the 10 transcripts with the highest expression level for downstream functional analysis. The results are shown in Table 2-Table 5.

Table 2. Expression statistics of Rhododendron fortunei transcripts

transcript_id	length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN3516_c0_g1_i1	2466	227980.54	5910.76	5288.32	98.52
TRINITY_DN6236_c1_g1_i3	6432	210543.33	1956.32	1750.31	76.43
TRINITY_DN481_c0_g1_i14	1136	162558.85	10545.91	9435.37	97.53
TRINITY_DN142_c0_g3_i4	2121	101636.24	3121.14	2792.47	82.05
TRINITY_DN117_c0_g1_i2	908	101297.15	8850.94	7918.89	99.79
TRINITY_DN211_c0_g1_i1	8742	98444.3	665.89	595.77	96.17
TRINITY_DN6137_c1_g1_i2	839	79074	7719.46	6906.57	100
TRINITY_DN1198_c0_g2_i1	1794	72136	2684.57	2401.88	100
TRINITY_DN54_c0_g2_i1	3133	68028	1355.59	1212.84	100

Table 3. Expression statistics of Rhododendron mariesii transcripts

transcript_id	length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN3994_c0_g1_i11	6806	763505.37	5592.43	4896.67	99.07
TRINITY_DN5067_c0_g1_i1	1497	359791.45	14120.74	12363.97	99.95
TRINITY_DN350_c1_g2_i2	7882	226127.08	1421.91	1245.01	85.01
TRINITY_DN3063_c0_g3_i3	1210	180616.97	9273.52	8119.8	94.82
TRINITY_DN25_c0_g3_i1	2355	110121	2535.84	2220.35	100
TRINITY_DN660_c0_g1_i1	950	86319.43	6148.82	5383.84	60.48
TRINITY_DN738_c0_g1_i3	1051	67964.61	4208.09	3684.56	49.78
TRINITY_DN2262_c0_g1_i1	2897	65459.45	1195.33	1046.62	99.18
TRINITY_DN883_c0_g1_i1	2429	64828	1441.46	1262.13	100

Table 4. Expression statistics of Rhododendron molle transcripts

transcript_id	length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN3994_c0_g1_i11	6806	763505.37	5592.43	4896.67	53.99
TRINITY_DN5067_c0_g1_i1	1497	359791.45	14120.74	12363.97	22.18
TRINITY_DN350_c1_g2_i2	7882	226127.08	1421.91	1245.01	66.64
TRINITY_DN3063_c0_g3_i3	1210	180616.97	9273.52	8119.8	60.55
TRINITY_DN25_c0_g3_i1	2355	110121	2535.84	2220.35	38.95
TRINITY_DN660_c0_g1_i1	950	86319.43	6148.82	5383.84	88.23
TRINITY_DN738_c0_g1_i3	1051	67964.61	4208.09	3684.56	33.78
TRINITY_DN2262_c0_g1_i1	2897	65459.45	1195.33	1046.62	100
TRINITY_DN883_c0_g1_i1	2429	64828	1441.46	1262.13	100

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

Table 5. Expression statistics of Rhododendron simsii transcripts

transcript_id	length	expected_count	TPM	FPKM	IsoPct
TRINITY_DN3994_c0_g1_i11	6806	763505.37	5592.43	4896.67	82.95
TRINITY_DN5067_c0_g1_i1	1497	359791.45	14120.74	12363.97	26.62
TRINITY_DN350_c1_g2_i2	7882	226127.08	1421.91	1245.01	71.74
TRINITY_DN3063_c0_g3_i3	1210	180616.97	9273.52	8119.8	100
TRINITY_DN25_c0_g3_i1	2355	110121	2535.84	2220.35	95.49
TRINITY_DN660_c0_g1_i1	950	86319.43	6148.82	5383.84	99.29
TRINITY_DN738_c0_g1_i3	1051	67964.61	4208.09	3684.56	88.67
TRINITY_DN2262_c0_g1_i1	2897	65459.45	1195.33	1046.62	34.84
TRINITY_DN883_c0_g1_i1	2429	64828	1441.46	1262.13	67.1

3.2. Functional Identification of Four Rhododendron Transcripts

We obtained the sequence information of four plant transcripts by splicing the transcriptome, and then used the software TransDecoder to predict the information of open reading frames in the transcriptome. Finally, the predicted sequences are aligned to the Swiss-prot and Pfam databases, and the results are shown in Table 6. Blastp and hmmscan tools were used for sequence alignment, and the results were used as input to predict the encoded protein sequence with TransDecoder software. In addition, we found that there are 8758 genes in four species of Rhododendron, 10472 genes in three species of Rhododendron, 12801 genes in two species of Rhododendron, and 20329 genes are unique to each species of Rhododendron.

For the functional annotation of four Rhododendron species, we use metascape software to annotate the go function, and the results are shown in Figure 5-Figure 8. It can be seen that the four rhododendrons are involved in the metabolism of RNA and non coding RNA.

Table 6. Functional notes of four Rhododendrons

SRA_ID	species	Swiss-prot	pfam
SRR5247112	Rhododendron simsii Planch.	16651	13256
SRR5247113	Rhododendron mariesii Hemsl.&E.H.Wilson	12344	11804
SRR5247114	Rhododendron molle (Blume) G. Don	11963	11375
SRR7686809	Rhododendron fortunei Lindl.	11402	11160

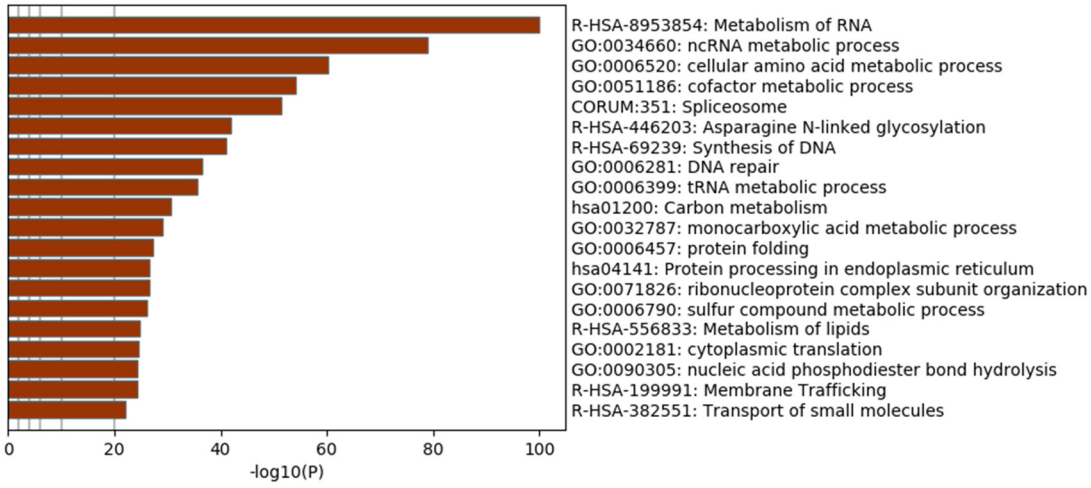

Figure 5. GO annotation of Rhododendron simsii transcripts

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

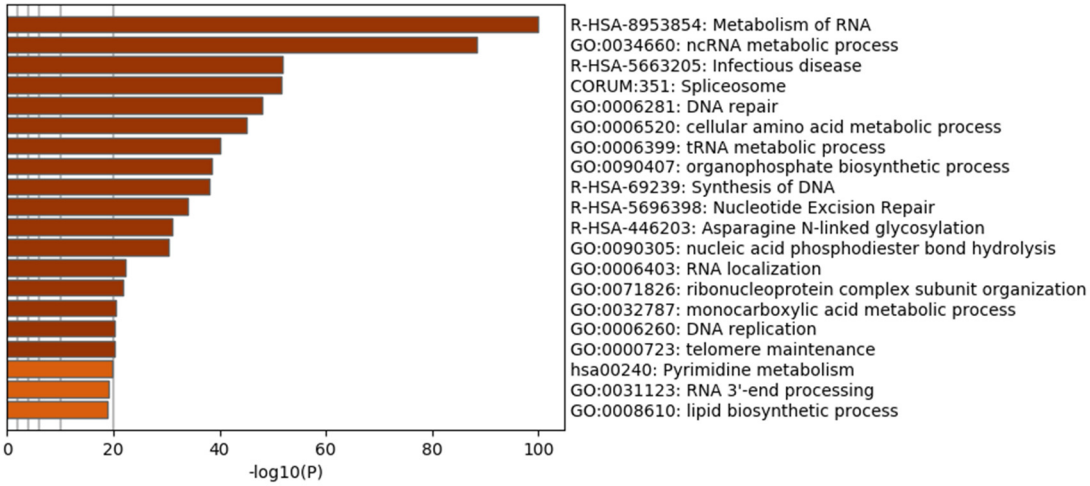

Figure 6. GO annotation of Rhododendron mariesii transcript

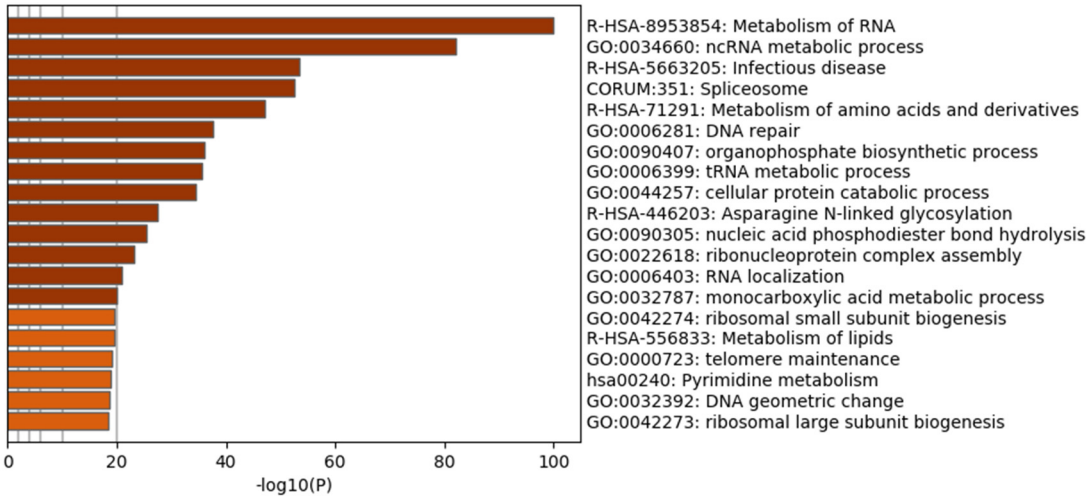

Figure 7. GO annotation of Rhododendron molle transcripts

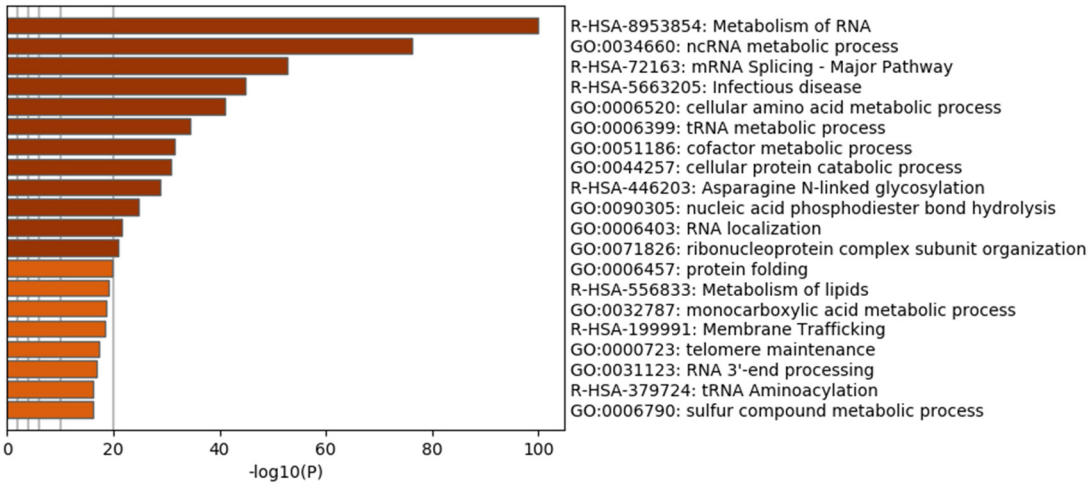

Figure 8. GO annotation of Rhododendron fortunei transcripts

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

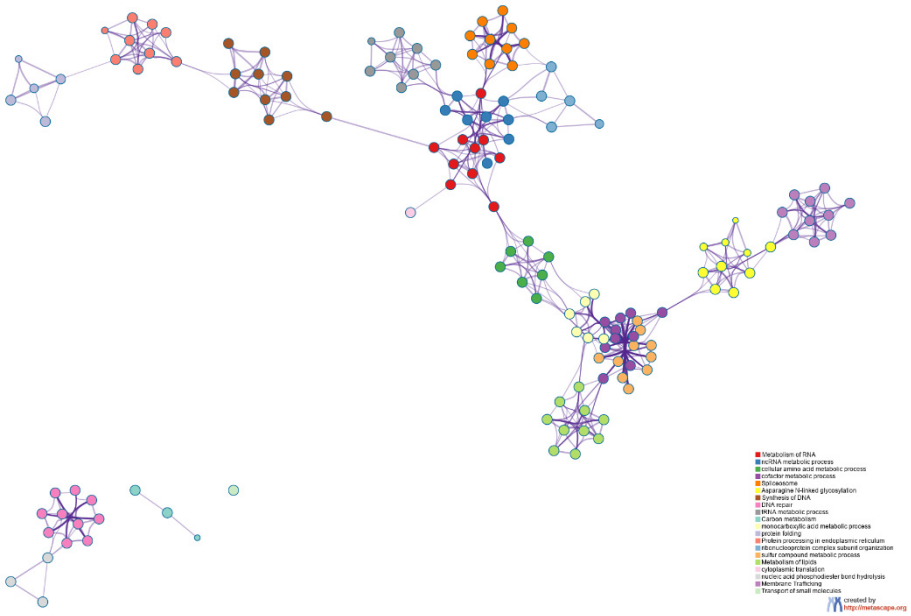

Figure 9. Enrichment pathway analysis of Rhododendron simsii

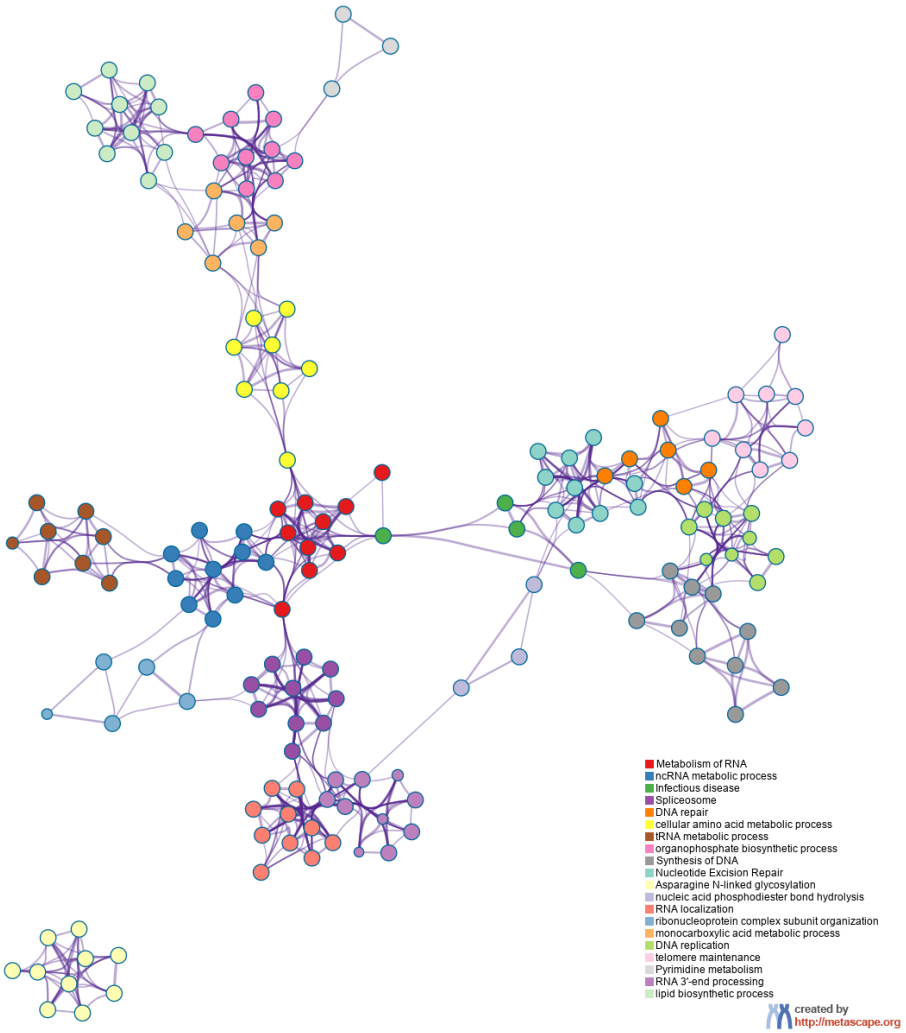

Figure 10. Enrichment pathway analysis of Rhododendron mariesii

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

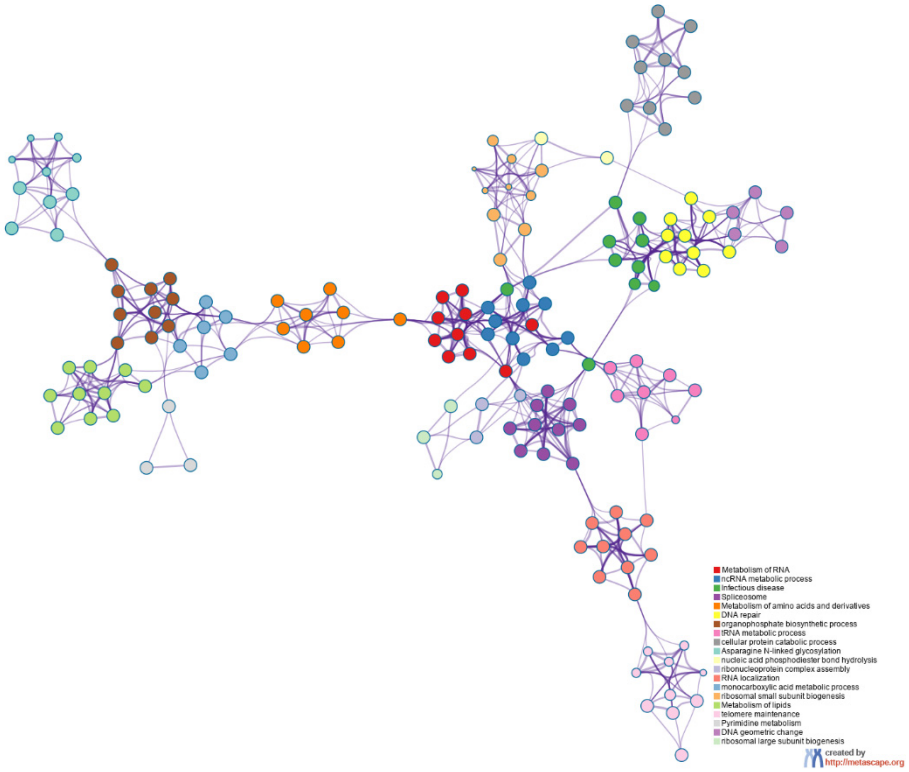

Figure 11. Enrichment pathway analysis of Rhododendron molle

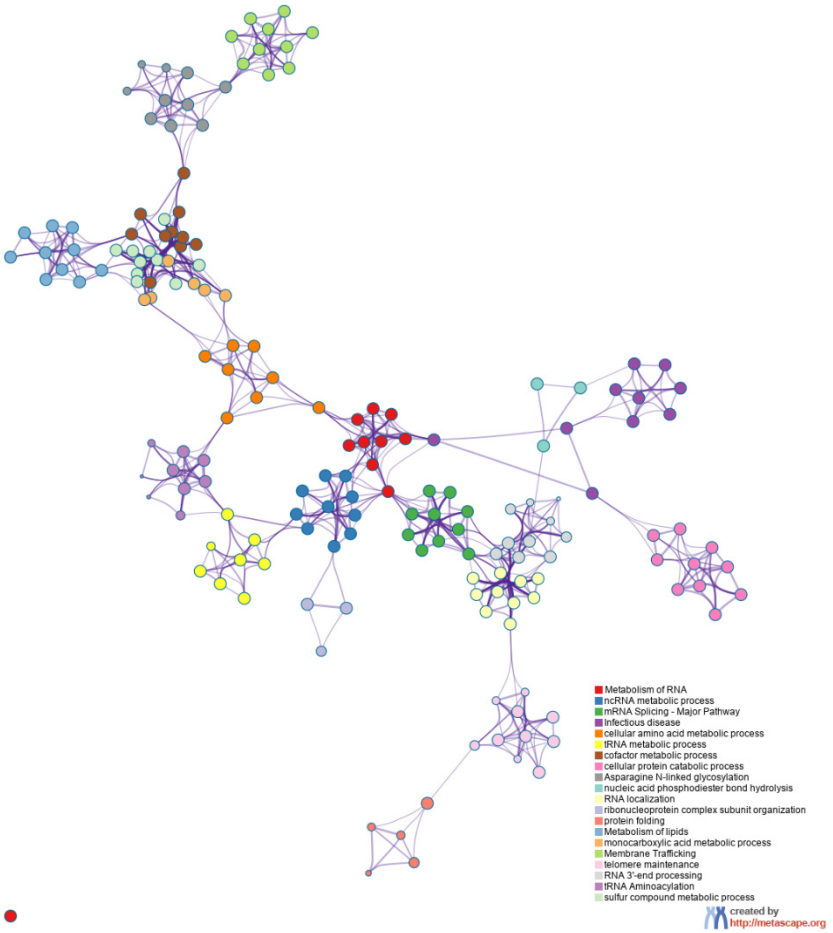

Figure 12. Enrichment pathway analysis of Rhododendron fortunei

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

3.3. Identification of Active Components in Four Rhododendron Species

According to the difference of transcripts expression among four Rhododendron species, we identified the top nine transcripts with the highest expression level, and obtained their gene functions through annotation.

Firstly, in Rhododendron fortunei, the transcripts TRINITY_DN3516_c0_g1_i1 encodes CYP92C6 gene, which is involved in the synthesis of volatile terpenoids, mainly converting geraniol to trimethyltridecanoic acid. Secondly, the transcripts trinity of Rhododendron molle flower tissue TRINITY_DN679_c0_g1_i1 encodes HMR2B, which catalyzes the synthesis of valproate, a specific precursor of all isoprene compounds in plants. Triterpenoid saponins (such as ginsenoside or ginsenoside) and phytosterol biosynthesis pathway promote the accumulation of triterpenoids in roots.

Through GC-MS detection of the tissue extract of Rhododendron molle, we obtained the active component maps identified by three extraction methods, and the results are shown in Figure 13. It can be seen from the figure that many active components can be separated by three extraction phases, such as 5-hydroxymethyl furfural, n-hexadecanoic acid, 9,12-octadecadienoic acid (Z, z) - and so on.

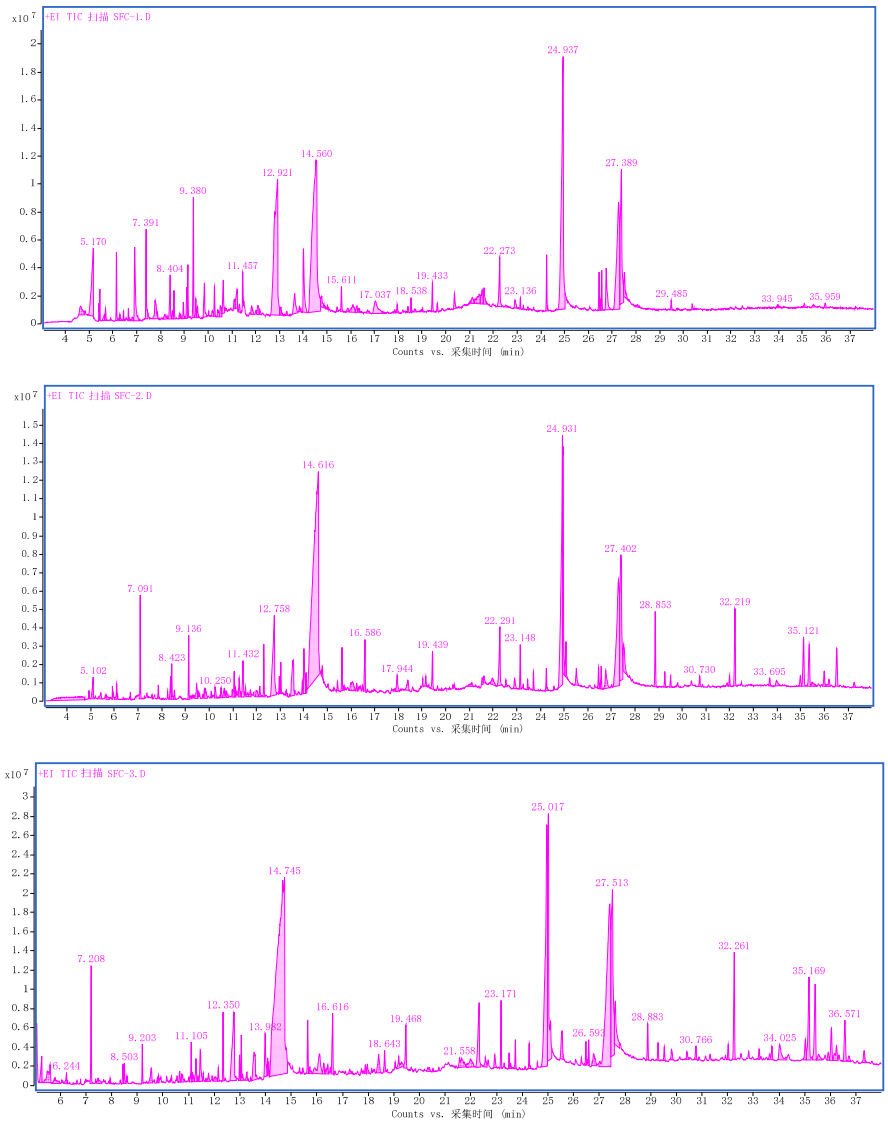

Figure 13. GC-MS chromatogram of Rhododendron molle flower tissue (methanol, ethanol, benzyl alcohol extraction method respectively)

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

4. Discussion

At present, there are not many reports on the omics of Rhododendron, and the published high-throughput sequencing data are also very few. By searching the SRA of NCBI database, there are only 321 pieces of Rhododendron's genomic data, most of which are related to soil microorganism's metagenomic sequencing data, and only a few of which are Rhododendron's transcriptome sequencing data. Therefore, we selected four kinds of Rhododendron's transcriptome data to study. Because the reference genome of Rhododendron has not been published, the lack of important information of gene annotation greatly limits the accuracy of research. Therefore, this study adopts the splicing strategy of no reference genome to process transcriptome data. This splicing method is widely used in no reference species, but also has the limitation of accurate splicing. Then, functional annotation analysis of the transcripts from four Rhododendron species was carried out to identify the gene expression sequences related to the active components. Finally, more than 80 metabolites were identified by GC-MS.

With the development of high-throughput sequencing technology, massive multi omics data have been produced. How to mine effective information from a large number of data has become the focus of research. In this paper, we used transcriptomic data to study the gene expression profiles of four Rhododendron species, and compared the expression differences of related species. The results showed that the four Rhododendron species shared 8758 homologous genes [17]. Gas chromatography-mass spectrometry was used to detect the flower tissue extract of Rhododendron molle, and the metabolites were identified to map to the differential expression of genes [18-19]. In the study of Rhododendron, due to the serious lack of molecular genetic data, and most of the studies focused on the adaptation of plant root nutrition and the identification of some SSR sequences [16]. Therefore, how to use the existing data resources to mine and study Rhododendron is the purpose and significance of this paper.

In some databases, the available research data of Rhododendron is very valuable, and the research direction is also very different. For example, Fang et al. Used transcriptome sequencing to study the transcriptional regulation process of photosynthesis in Rhododendron, in which they mainly identified heat responsive genes and focused on the response of plants to stress [15]. Zhou et al. Extracted and separated the effective components from the flower tissue of Rhododendron molle, and then identified the diterpenoids by crystallization technology combined with NMR to obtain the complete active substance configuration [11-14]. Liu et al. Studied the content analysis of anthocyanins and flavonols in petals of 10 species of Rhododendron in southeastern Taipei, and identified the active components by HPLC-MS [7]. Lai et al. Mainly studied the neuroprotective activity of 6,8-di-c-methyl-flavonoids of Rhododendron fortunei. By separating the twigs and leaves of Rhododendron fortunei, the extracts were analyzed by spectral and chemical methods, and the neuroprotective effect of the extracts was finally verified by cell experiment [5]. Ahmed Rezk et al. Studied the antibacterial activity of leaf tissue extracts from Rhododendron plants, and selected 26 species of bacteria from different taxonomic branches. Agar diffusion test was carried out with 80% methanol extracts of 120 Rhododendron species. Principal component analysis was used for data analysis [3].

In this paper, transcriptome sequencing data mining method was used to study the active components of four Rhododendron species, and GC-MS was used to verify the components of Rhododendron species. Transcriptome sequencing data can be used to analyze gene expression profile information, identify gene differential expression, and construct expression regulatory network through gene annotation; Finally, GC-MS technology is used to verify the accuracy of data mining information, which makes the research more systematic and reasonable.

ISSN: 2710-0170 DOI: 10.29561/FHSS.202106_1(3).0002

References

[1] Song Hejiao. Chemical constituents of two non-toxic medicinal plants of Rhododendron . Kunming University of science and technology, 2009.

[2] Zhou Junfei. Chemical constituents and bioactivity of Rhododendron molle leaves.

[3] Rezk A , Nolzen J , Schepker H , et al. Phylogenetic spectrum and analysis of antibacterial activities of leaf extracts from plants of the genus Rhododendron. BMC Complementary and Alternative Medicine, Vol.15(2015) No. 1, p.67.

[4] Xiangying W , Jianjun C , Chunying Z , et al. A New Oidiodendron maius Strain Isolated from Rhododendron fortunei and its Effects on Nitrogen Uptake and Plant Growth. Frontiers in Microbiology, (2016), p.7.

[5] Lai Y, Zeng H , He M , et al. 6,8-Di-C-methyl-flavonoids with neuroprotective activities from Rhododendron fortunei. Fitoterapia, Vol.112 (2015), p.237-243.

[6] Li Y , Liu Y B , Zhang J J , et al. Antinociceptive Grayanoids from the Roots of\r, Rhododendron molle. Journal of Natural Products, (2015)acs.jnatprod.5b00456.

[7] Liu L , Zhang L Y , Wang S L , et al. Analysis of anthocyanins and flavonols in petals of 10 Rhododendron species from the Sygera Mountains in Southeast Tibet. Plant Physiology and Biochemistry, Vol.104 (2015) p.250-256.

[8] Yong-Qing C , Jian-Hui H U , Jie Q , et al. Rhododendron Molle(Ericaceae): phytochemistry, pharmacology, and toxicology. Chinese natural medicine, (2018).

[9] Li Y , Liu Y B , Yan H M , et al. Rhodomollins A and B, two Diterpenoids with an Unprecedented Backbone from the Fruits of Rhododendron molle. Scientific Reports, Vol. 6(2016), p.36752.

[10] Xiao Z , Su J , Sun X , et al. De novo transcriptome analysis of Rhododendron molle G. Don flowers by Illumina sequencing. Genes & Genomics, (2018).

[11] Zhou, Shuai-Zhen, Yao, Sheng, Tang, Chunping. Diterpenoids from the Flowers of\\r, Rhododendron molle. Journal of Natural Products, Vol.77(2014)No. 5, p.1185-1192.

[12] Zhou J , Sun N , Zhang H , et al. Rhodomollacetals A–C, PTP1B Inhibitory Diterpenoids with a 2,3:5,6-Di-\r, seco\r, -grayanane Skeleton from the Leaves of\r, Rhododendron molle. Organic Letters, (2017) acs.orglett.7b02633.

[13] Zhou J , Liu T , Zhang H , et al. Anti-inflammatory Grayanane Diterpenoids from the Leaves of Rhododendron molle. Journal of Natural Products, Vol. 81(2017)No.1, p.151.

[14] Zhou J , Zhan G , Zhang H , et al. Rhodomollanol A, a Highly Oxygenated Diterpenoid with a 5/7/5/5 Tetracyclic Carbon Skeleton from the Leaves of\r, Rhododendron molle. Organic Letters, Vol. 19 (2017) No.14, p.3935-3938.

[15] Fang L , Tong J , Dong Y , et al. De novo RNA sequencing transcriptome of Rhododendron obtusum identified the early heat response genes involved in the transcriptional regulation of photosynthesis. Plos One, Vol. 12(2017)No.10, p.e0186376.

[16] Yue Z , Xue Z , Yue-Hua W , et al. De Novo Assembly of Transcriptome and Development of Novel EST-SSR Markers in Rhododendron rex Lévl. through Illumina Sequencing. Frontiers in Plant Science, Vol. 8(2017),p.1664-.

[17] Xiangying W , Jianjun C , Chunying Z , et al. Differential Gene Expression in Rhododendron fortunei Roots Colonized by an Ericoid Mycorrhizal Fungus and Increased Nitrogen Absorption and Plant Growth. Frontiers in Plant Science, (2016),p. 7.

[18] Qian Chen-Yu, Quan Wen-Xuan, Xiang Zhang-Min, Li, Chao-Chan, Characterization of Volatile Compounds in Four Different Rhododendron Flowers by GC×GC-QTOFMS. Molecules, Vol.24 (2019), p. 3327.

[19] Vandana G , Sukhmeen K , Saroj A , et al. Antioxidant and Antimutagenic Activities of Different Fractions from the Leaves of Rhododendron arboreum Sm. and Their GC-MS Profiling. Molecules, Vol. 23(2018) No.9, p.2239-.