Illumina-Based Transcriptome Analysis of Four Genotypes of Pinellia ternata (Thunb.) Breit.

Yong Jing
Chengdu University of Traditional Chinese Medicine
Yueyue Lai
Chengdu University of Traditional Chinese Medicine
Ziyu Wan
Chengdu University of Traditional Chinese Medicine
Qiao Li
Chengdu University of Traditional Chinese Medicine
Min Li (✉ 19862025@cdutcm.edu.cn)
Chengdu University of Traditional Chinese Medicine
Lijuan Liu
University of Macau
Zhejie Chen
University of Macau
Shengpeng Wang
University of Macau
Yitao Wang
University of Macau
Xiao Huang
Chengdu University of Traditional Chinese Medicine

Research

Keywords: Pinellia ternate, Transcriptome, Functional annotation, Unigenes, KEGG Classification

Posted Date: April 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-18919/v1

Abstract

Background Pinellia ternate (Thunb.) Breit. (P. ternate), a weed species distributed in Eastern Asia, has been widely used as a Chinese herbal medicine. Currently, four genotypes of P. ternate have been identified according to their different leaf shapes, including Shaoye, Taoye, Zhuye and Liuye Type. In addition, the total alkaloids and organic acids content in tuber, tuber yields and other physiology and agronomic characteristics were significant difference. However, little is known about the evolution and genetic diversity of four genotypes of P. ternate . Moreover, the genomics of P. ternate remains uncharacterized. The recent development of RNA-Seq, a next generation sequencing technology, provides an opportunity to expand the identification of P. ternate genes through in-depth transcript profiling. Methods In this study, we presented a comprehensive landscape of the transcriptome profiles of four genotypes of P. ternate using Illumina paired-end sequencing technology. Results Totally, 70,491,871, 72,819,035, 73,159,913 and 71,612,115 reads were achieved in four genotypes of P. ternate , respectively. A total number of 126,391 assembled unigenes were identified. After aligning the sequences to the public protein databases, 82,998 unigenes were annotated. Among these, 15,534 were assigned to Gene Ontology categories, 41,216 to Clusters of Orthologous Groups, and 46,053 to Kyoto Encyclopedia of Genes and Genomes pathways. In addition, 23,678 simple sequence repeats (SSRs) and 65,527 Single Nucleotide Polymorphism (SNP) variations were detected. Conclusion This study provides abundant genomic data and comprehensive transcriptome for P. ternate . We envision that these transcriptome datasets will serve as an important public information platform to accelerate further studies of the P. ternate and the identification of genes that potentially involving in the biosynthesis of total alkaloids, acid and other active ingredients among different genotypes P. ternate .

Background

Pinelliae Rhizoma (Banxia in Chinese), a famous Chinese herbal medicine derived from the dried tuber of Pinellia ternata (Thunb.) Breit. (P. ternate), is widely prescribed for the treatment of cough, phlegm, headache, and vomiting in clinic (Fig. 1) [1]. Containing a variety of chemical constituents, including organic acids, amino acids, alkaloids and etc., Pinelliae Rhizoma has shown multiple biological activities, such as anti-tumor, anti-inflammation, anti-bacteria, and anti-pregnancy activities, and has been reported to have significant impacts on respiratory, digestive, cardiovascular, and nervous systems [2, 3]. As a weed and cultivated Chinese herbal medicine, P. ternate widely disperse in China, including Sichuan, Hubei, Gansu, Yunnan, Guizhou, Henan and Anhui Provinces. Among these regions, Sichuan, Gansu and Yunnan Provinces are the most suitable areas for cultivation of P. ternate. Generally, P. ternate can be classified into four types based on their different shaped leaves according to the different leaf length/width ratios, namely Shaoye, Taoye, Zhuye and Liuye Type, which are shortened to SK, TK, ZK and LK, respectively. Notably, their distinct total alkaloids and organic acids contents in tuber, tuber yields, and other physiology and agronomic characteristics are also the factors to distinguish one of them from another. Generally, the difference of chemical composition of four types of P. ternate potentially leads to the inconsistency of drug intervention effect in clinical, and the physiological discrepancy of plant directly affect the unified management in the process of herb planting. Therefore, the use of modern technology, especially gene technology to effectively distinguish the four types of P. ternate is of great significance to the modernization of the cultivation and scientific clinical applications of this plant. Regretfully, the studies about the DNA or protein sequences of four genotypes of P. ternate were limited.

In recent years, researchers have focused on the investigation of the genetic diversity of P. ternate. For example, Chen et al. reported 12 microsatellite markers for P. ternate and developed the markers as useful tools for studying the genetic diversity of the Pinellia species [4]. The group of Chung authenticated several Chinese and Korean herbal medicines by RAPD analyses and PCR-RFLP [5]. Moreover, 54 miRNAs from 23 miRNA families have been identified by Xu. et al [6]. To date, however, DNA sequence information of P. ternate and its encoded genes has not been well-documented. Sequencing of large genomes is highly time and cost consuming, thus transcriptome analysis is an alternative and cost-effective approach for discovering new genes, providing information on gene expression and regulation, as well as the amino acid contents of proteins [7–9]. Herein, we survey the poly (A) + transcriptome of P. ternate tissues by employing Next-Generation Sequencing (NGS) technology. A total of 35.46 G raw reads was collected, offering comprehensive information of all known transcriptomes and the major metabolic pathways. The dataset contains assembled and annotated transcriptome sequences could serve as a public information platform providing gene expression, genomics, and functional genomics of P. ternate and further stimulating the identifications of more transcriptome sequences.

Materials And Methods

The Minimum Standards of Reporting Checklist contains details of the experimental design, and statistics, and resources used in this study (Additional file 1).

Sample preparation

Four different genotypes of *P. ternate* were cultivated in the Chinese herbal medicine standardized planting base under Good agricultural practices (GAP) in Hongjiang town, where was located in Pengxi County of Sichuan Province, China (Fig. 2). Routinely, the four genotypes of *P. ternate* were collected from three parallel samples in September, when the plants were in the vigorously growing stage. After collection, the tubers were cut into small pieces and immediately frozen in liquid nitrogen. All samples were stored at -80°C for further analysis.

Determination of total organic acids

An aliquot of 5.0 g sample powder (through 250 mesh screen) was accurately weighed and extracted by refluxing with 50 mL of ethanol for three times. The combined extracts were evaporated to dryness under a stream of nitrogen. The residue was suspended in 10 mL of 0.1 M NaOH solution and extracted by sonication for 30 min. Then, the extract solution was diluted with water to a constant volume (50 mL), and 25 mL was used for the determination of total organic acids using potentiometric titration.

Determination of total alkaloids

An aliquot of 0.5 g sample powder (through 250 mesh screen) was accurately weighed and extracted by cold soaking for 10 h after sequential addition of 0.5 mL of ammonium hydroxide ethanol and 10 mL of chloroform. The extract was then extracted by sonication for 45 min and evaporated to dryness under a stream of nitrogen. The chemical standard of ephedrine hydrochloride was dissolved in water to a desired concentration of 0.2 mg/mL as the stock solution. The calibration curves were prepared by diluting appropriate volumes of stock solutions into seven different concentrations according to the following procedures: 1) Appropriate volume of stock solutions was diluted with water to a constant volume (2 mL); 2) After 8 mL of chloroform, 10 mL of citric acid-sodium citrate buffer and 1 mL of bromothymol blue standard solution were added, the mixed solution was shaken for 15 min and then chloro- form extract was collected; 3) The chloro- form extract was diluted with chloroform to 20 mL as working solution. All working solutions were measured as 418 nm and the calibration curve constructed by plotting UV absorbance (y) *vs.* the corresponding concentrations (x). was calculated as Y = 27.679X + 0.0654 with good linearity (R^2=0.9991).

RNA Extraction

Total RNA was extracted from these samples using TRIzol (Thermo Scientific, MA, USA) according to the manufacturer’s instructions. The extracted RNA was treated with RNase-free DNase I (New England BioLabs) for 30 min at 37°C in order to remove residual DNA. Equal amount of the extracted RNA from each sample were prepared for further experiments.

mRNA-seq Library Construction for Hiseq 2000 Sequencing

Extracted RNA samples were purified by oligo-dT beads (Dynabeads mRNA purification kit, Invitrogen), then polyA-containing mRNA were fragmented into 200-250bp with Fragment buffer (Ambion, TX, USA). First strand cDNA was synthesized in the mixture buffer containing N6 primer, First Strand Master Mix and Super Script II reverse transcription (Thermo Scientific, MA, USA) with the following reaction program: 10 min at 25℃, 30 min at 42℃, followed by 15 min at 70℃, and eventually hold at 4℃. Subsequently, Second Strand Master Mix was added and the double-stranded cDNA were generated in the reaction system for 2 h at 16℃. After purification using QIAquick PCR Purification Kit (QIAGEN, CA, USA), the End Repair Mix was added to convert the overhangs of the double-stranded cDNA into blunt ends and the mixture were incubated for 30 min at 20℃. Then, the purified DNA was mixed with A-Tailing Mix buffer before being incubated at 37℃ for 30 min. After purification and elution, the DNA samples were incubated with the single ‘A’ nucleotide in a thermal cycler for 30 min at 37℃ to protect DNA from ligating to one another during the adapter ligation reaction with Adapter and Ligation Mix. The adapters were ligated to the ends of the DNA fragments at room temperature for 20 min. The products of the ligation reaction were purified on a 2% agarose gel to remove unbounded adapters and adapters ligated to one another. A narrow size-range of 300-350 bp of sequencing library was selected and purified with QIAquick Gel Extraction kit (QIAGEN). Then, the enrichment of cDNA fragments was performed using PCR and finally purified with Ampure XP Beads (Agencourt, MA, USA). The library was validated by determining the average molecule length using the Agilent 2100 bioanalyzer instrument (Agilent DNA 1000 Reagents) and measuring the DNA concentration by real-time quantitative PCR (qPCR) (TaqMan Probe). The qualified library was clustered on the flowcell using an Illumina cBot and the amplified flowcell was sequenced using paired-end sequencing technology on the HiSeq 2000 System (TruSeq SBS KIT-HS V3, Illumina) with the most common sequencing strategy of a read length of 90. Data analysis and base calling were carried out using Illumina instrument software.

Sequence Data Assessment and Assembly

Clean reads were trimmed with the removing of reads in which the number of quality bases contained adaptor sequences, reads in which number of quality bases were more than 10% of the N-base, and reads in which more than half of the quality values of the bases were less than 5. The clean reads were then calculated with CycleQ 20 and the qualified reads were assembled into contigs by overlapping between the sequences using Trinity software. According to paired-end information of the sequences, the contigs were joined into transcript sequences to form isotigs, recovering full-length transcripts across a broad range of expression levels and displaying a similar sensitivity compared to the general genome alignments methods [10]. In this assembly, the overlap settings were 24 bp and 80% similarity, along with group pairs distance was set to maximum length 500 and all the other parameters were set to the default values. The longest transcriptions from the assembled component alternative splicing transcripts were chosen as unigene sequences. The transcript levels in Fragments per kilobase of exon mode per million mapped reads (FPKM) were quantified [11]. The measurement of FPKM of fragment density indicated the molar concentration of a transcript in the starting sample by normalizing for RNA length and the total fragment number.

Unigenes Annotation

The predicated amino acid sequences encoded by the assembled unigene sequences were annotated using the National Center for Biotechnology Information (NCBI) non-redundant protein (Nr) database and Swiss-Prot database. Using BlASTx (version 2.2.14), the cutoff criteria for E-value of annotation was set to $<10^{-5}$. Based on the best BLAST hit (highest score), gene names were given and searches were limited to the first 10 significant hits for each query. EMBOSS software package was utilized for predicting the open reading frames (ORFs) for each unigene through the “getorf” program [12].

To further illustrate gene ontology (GO) terms of assembled unigenes, including molecular functions, biological processes and cellular components, the NCBI BlAST results were imported into Blast2GO software [13, 14]. These GO terms were assigned to query sequences. A broad overview of groups of genes catalogued in the transcriptome for molecular functions, biological processes and cellular components were produced and further enriched and refined by ANNEX [15, 16].

Moreover, the Clusters of Orthologous Group (COG) database were employed to predict and classify functions of unigene sequences [17]. By using the online KEGG Automatic Annotation Server (http://www.genome.jp/kegg/kaas/) and the bi-directional best hit method, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were assigned to the assembled unigenes and KEGG Orthology (KO) assignment was obtained [18]. The output of KEGG analysis contained KO assignments and KEGG pathways that were populated with the KO assignments.

EST-simple sequence repeat (SSR) Detection

The online program SSR Identification Tool (http://www.gramene.org/db/markers/ssrtool) was employed to identify SSRs in a given sequence [19, 20]. The parameters were set to identify perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of six, five, four, four, and four repeats, respectively. The information of the online report listed the total number of sequences containing SSRs among the submitted unigenes, sequence ID, SSR motifs, number of repeats (di-, tri-, tetra-, penta-, and hexanucleotide repeat units), repeat length, SSR starts, and SSR ends [21-23].

Statistical analysis

The experimental data were obtained from 3 independent tests and statistical analysis was performed by one-way analysis of variance (ANOVA) using SPSS Statistics (version 24.0, International Business Machines Corp, NY). *P* < 0.05 was considered as a statistically significant difference.

Results

Analysis of Total Alkaloids and Organic Acids in Tubers of Four Genotypes of P. ternate

Preliminarily, we analyzed the total alkaloids and organic acids contents in tuber from four genotypes of *P. ternate* (SK, TK, ZK and LK). ZK was reported to have a preponderance of total alkaloids (0.047%) and organic acids (0.556%), while LK was shown the lowest content of total alkaloids (0.027%) and organic acids (0.363%) (Table 1).

Table 1 Total alkaloids and organic acids in tubers of four genotypes of *P. ternate* (n=3)

Samples	Total alkaloids (%)	Total organic acids (%)
SK	0.0438	0.437
TK	0.0405	0.415
ZK	0.0487	0.556
LK	0.0270	0.363

Transcriptome Sequencing and De Novo Assembly of Four Genotypes of P. ternate

To better understand the genotypes of SK, TK, ZK and LK associated with formation and accumulation of chemical compounds, we performed RNA-seq. RNA extracted from the tubers of the four genotypes were used to create cDNA libraries. Totally, 70,491,871, 72,819,035, 73,159,913 and 71,612,115 reads were collected from the SK, TK, ZK and LK cDNA libraries, respectively (Table 2). To ensure the reliability of the libraries, we performed quality controls and 64,911,333, 66,839,740, 66,217,035 and 65,469,304 clean reads were obtained for SK, TK, ZK and LK cDNA libraries, respectively. Using Trinity (version: v2.0.6) software, which has been demonstrated to be efficient for de novo reconstruction of transcriptomes from RNA-Seq data [10, 24], de novo assembly was then applied to construct transcripts from these RNA-seq reads. Furthermore, Tgicl (version: v2.0.6) program was performed to cluster transcripts to unigenes. Finally, a total number of 126,391 unigenes were obtained, together with a mean unigene length of 835 bp and an N50 value of assemblies of 1473 bp (Table 3). The sequence sizes distribution of unigenes was homogeneous (Fig. 3). Among these, 33,812 unigenes (26.75%) were longer than 1 kb while more than 57,781 unigenes (45.72%) were greater than 500 bp.

Table 2 Assessment of assembly quality of four genotypes of *P. ternate*

	SK	TK	ZK	LK
Raw reads	70491871	72819035	73159913	71312115
Clean reads	64911333	66839740	66217035	65469304

Table 3 Summary of unigenes number of length distribution

Length (bp)	Total Number
200-500	68610
500-1000	23969
1000+	33812
Total number of Unigenes	126391
Mean length of Unigenes	835
N50 length of Unigenes	1473

Functional Annotation

Unigene annotations provided functional information of query sequences, including protein sequence similarities, COG clusters, GO and KEGG pathway information. We annotated the *P. ternate* unigene sequences by aligning them with those deposited in protein databases (NCBI Nr, Swiss-Prot, KEGG and COG) using BLASTX with E-value >10^{-5}. There were 70,482 unigenes matched in the NCBI Nr database. Among them, 49,542 unigenes were significantly matched in the NCBI Nr database and 47,394 were similar to proteins that deposited in the Swiss-Prot database. A total number of 82,998 unigenes, about 65.67% of all assembled unigenes, were annotated in one or more of the databases (Table 4), suggesting assembled unigenes had relatively well conserved functions.

Table 4 Summary of functional annotations for *P. ternate* unigenes in public databases

Annotated databases	Number of unigenes	Percentage (%)
NR	70482	55.77
NT	49542	39.20
Swiss-Prot	47394	37.50
KEGG	46053	36.44
COG	41216	32.61
Interpro	54602	43.20
GO	15534	12.29
All databases	82998	65.67
Total unigenes	126391	

GO Classification

GO is an international standardized gene-function classification system that uses a dynamically updated, controlled vocabulary and a strictly defined concept to comprehensively describe the properties of genes and their products in any organism. The GO database comprises three ontologies, including molecular function, cellular components and biological processes. In this study, we obtained the GO functional annotations of the *P. ternate* unigenes by BLAST2 GO program [13]. Using the WEGO software [16], we performed GO functional classifications for all of the unigenes and examined the macro level distribution of gene functions of this species. Based on sequence homology, 15,534 unigenes were annotated in the GO database and categorized into the three main GO categories and 55 functional groups (Fig. 4). In the “biological process” category, the unigenes related to “cellular process” (49.33%) and “metabolic process” (54.08%) were predominant. The “cell” (51.07%) and “cell part” (51.07%) were found to be the most abundant classes. Under the “molecular function” category, the majority of unigenes were involved in “binding” (47.03%) and “catalytic activity” (51.30%).

KEGG Classification

KEGG is well known as providing a basic platform for systematic analysis of gene functions in terms of the networks of gene products [25]. To further identify the biological pathways that are active in *P. ternate*, the analysis of unigenes by KEGG showed that 46,053 unigenes were mapped to 127 reference canonical pathways (Table 4). These pathways were divided into five main categories: “cellular processes”, “environmental information processing”, “genetic information processing”, “metabolism” and “organismal systems”. Among these, 13,273 (28.82%) were mapped to metabolic pathways and 5,463 (11.86%) were associated with biosynthesis of secondary metabolites. The major pathways involving thousands of unigenes were RNA transport (ko03013) (5,024 unigenes, accounting for 10.91%) , mRNA surveillance pathway (ko03015) (3,882, 8.43%) , endocytosis (ko04144) (3,125, 6.79%) , glycerophospholipid metabolism (ko00564) (3,017, 6.55%), ether lipid metabolism (ko00565) (2,718, 5.9%) , spliceosome (ko03040) (2,186, 4.75%) , plant-pathogen interaction (ko04626) (1,781, 3.87%) , ribosome (ko03010)

(1,737, 3.77%) , purine metabolism (ko00230) (1,574, 3.42%) , plant hormone signal transduction (ko04075) (1,507, 3.27%) (Fig. 5).

SSR and SNP Detection

Molecular markers play important roles in studying molecular biology (e.g., gene mapping) and molecular breeding [26, 27]. Herein, we aimed to identify the potential molecular markers for the molecular research community and breeding of *P. ternate*. Two types of markers, SSR (Simple Sequence Repeats) and SNP (Single Nucleotide Polymorphism) were investigated. MISA (version: v1.0) (http://pgrc.ipk-gatersleben.de/misa) [28] was employed to indentify SSR in Unigenes and primers for each SSR were design by Primer3 (version: v2.2.2) (http://bioinfo.ut.ee/primer3) [29], We mapped all clean reads of Unigenes using HISAT (version: v0.1.6-beta) (http://ccb.jhu.edu/software/hisat/index.shtml) [30] and used GATK with (version: v3.4-0) (https://www.broadinstitute.org/gatk) [31] pipeline for SNP calling. Totally, 23,678 SSRs were identified from 126,391 unigenes, accounting for 18.73% of all unigenes. Additionally, 4,133 unigenes contained more than 1 SSR. SSRs generally included 2 to 6 nucleotide repeat types and the number of repeats changed significantly among genotypes. The percentages of mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeat SSRs of all the SSRs in this study were about 11.58%, 44.56%, 39.59%, 14.78%, 11.66% and 16.22%, respectively (Table 5). On the other hand, 65,527 SNP variations were also identified from the four genotypes. The high density SNP markers could be applied as useful tools for molecular research of *P. ternate* in case of SSR markers are not available.

Table 5 Summary of SSR types in *P. ternate*

Searching items	Numbers	Percentage (%)
Total number of sequences examined	126391	-
Total number of identified SSRs	23678	-
Number of sequences containing more than one SSR	4133	-
Number of SSRs present in compound formation	2243	-
Mono-nucleotide	2742	11.58
Di-nucleotide	10552	44.56
Tri-nucleotide	9374	39.59
Tetra-nucleotide	350	14.78
Penta-nucleotide	276	11.66
Hexa-nucleotide	384	16.22

Differential Expression of Unigenes of Four Genotypes of *P. ternate*

The four genotypes (SK, TK, ZK and LK) with different leaf shapes and different amounts of total alkaloids and organic acids in their tubers showed relative discrepancy of their transcript. Consequently, we sought to analyze the differential expression of unigenes in the four genotypes of *P. ternate*. The

NOIseq and PossionDis program were employed to identify the differential expression of unigenes among SK, TK, ZK and LK [32, 33]. For SK *vs.* TK, 4,255 unigenes were differentially expressed in which 1,883 unigenes were obviously up-regulated in TK whereas 2,372 unigenes were significantly down-regulated. Besides, for SK *vs.* ZK, 4,540 unigenes were differentially expressed, with 689 unigenes up-regulated in SK and 3,851 unigenes declined in ZK. Compared the unigenes expression of SK against that of LK, 1,386 unigenes in SK were increased while 4,391 unigenes in LK down-regulated. For TK *vs.* ZK, 1,987 unigenes were differentially expressed, with 777 unigenes promoted in TK and 1210 unigenes descended in ZK. For TK *vs.* LK, 1,100 unigenes were differentially expressed, in which 326 unigenes were up-regulated in TK and 774 unigenes were decreased in LK. And for ZK *vs.* LK, 1,828 unigenes were differentially expressed, with 340 increased unigenes in ZK and 1488 down-regulated unigenes in LK (Table 6).

Table 6 Summary of different expression of unigenes among SK, TK, LK and LK

	Up-regulated unigenes	Down-regulated unigenes	Total
SK vs TK	1386	4391	5777
SK vs ZK	1883	2372	4255
SK vs LK	326	774	1100
TK vs ZK	340	1488	1828
TK vs LK	689	3851	4540
ZK vs LK	777	1210	1987

To further analyze the possible function of unigenes with differential expression levels, we assessed their GO classifications. 1,206 unigenes with differential expression between SK and TK were classified into 110 pathways by KEGG analysis. Obviously, 509 unigenes were enriched in metabolic pathways as determined by KEGG analysis. The unigenes of SK *vs.* ZK, SK *vs.* LK, TK *vs.* ZK, TK *vs.* LK and ZK *vs.* LK were also enriched in metabolic pathways by KEGG analysis (Table 7). The occurrence of unigenes with differential expressions among samples suggested that other differences may contribute to the relatively larger number of differentially expressed unigenes.

Table 7 Summary of differential expression of unigenes among SK, TK, LK and LK as classified by KEGG analysis.

	Differential expression of unigenes	Pathways	Unigenes enriched in metabolic pathways
SK *vs* TK	1206	110	509
SK *vs* ZK	1341	110	605
SK *vs* LK	1434	118	790
TK *vs* ZK	707	106	415
TK *vs* LK	262	94	163
ZK *vs* LK	622	105	421

Discussion

The medication history of P. ternate in TCM has been more than 2,000 years, its phytochemistry and biological activities have been drawn the attention of researchers to constantly probe and further explore on it. The lack of genetic resources and genomic information has hampered the in-depth understanding of P. ternate in relationship between growth and medicinal properties. Herein, we first report the most comprehensive transcriptome analysis of four genotypes of P. ternate (Shaoye, Taoye, Zhuye and Liuye Type) using the Illumina platform. In our transcriptome analysis, the obtained total clean reads nucleotides, length distribution pattern and unigene lengths were similar to the previous Illumina based sequencing studies [34], indicating an effective de novo assembly of four genotypes of P. ternate transcriptome sequencing data. Among the 126,391 identified unigenes in transcriptome analysis, 46,053 (36.44%) unigenes were mapped to reference pathways and the rest (63.56%) still remains unknown.

GO functional classification is utilized to distribute gene functions at the macro level and predict the physiological role of each unigene [35, 36]. In our studies, 15,534 unigenes were annotated and categorized into the three main GO categories and 55 functional groups. The annotated results revealed that the metabolic pathways the assembled unigenes involved are as diverse as the molecular functions they served. Among the biological process categorized by GO classification, metabolic process and cellular process emerged as the predominant positions, indicating the occurrence of important cellular and metabolic activities in P. ternate. For the molecular function category, catalytic activity and binding were the most highly represented GO terms, which are the similar results to reported studies [37, 38]. The most abundant classes, cell and cell part, could not be matched to the previous reports, suggesting there is a limited information of genes functions of P. ternate.

KEGG pathways can help us better understand biological functions and gene interactions. In functional classification by KEGG, “metabolic pathways” represented the largest category, suggesting P. ternate invests in energy homeostasis and cell maintenance and defense. Unigenes associated with biosynthesis of other secondary metabolites were identified as the second largest group in the present study, with evidence that numerous biologically active secondary metabolites have been isolated and verified in P. ternate [39].

In SSR detection, dinucleotide repeats were the maximum type of SSR motifs in the transcriptome of P. ternate, while the tri-type motifs were second. Besides, more than two thousand SSRs were displayed in the compound formation, meaning DNA stretches consist of two or more different repeats. Furthermore, the analysis of high density of SNP form four genotypes of P. ternate can provide theoretical basis for species identification in the future.

Conclusion

Taken together, we performed an Illumina-based transcriptome study of four genotypes of *P. ternate*. The transcriptome analysis identified 126,391 unigenes, among which 46,053 (36.44%) unigenes were mapped to 127 reference canonical pathways. A total number of 23,678 SSRs and 65,527 SNPs were also validated, which may be useful for the further molecular research of *P. ternate*. A total number of

4,255, 4,550, 5,777, 1,987, 1,100 and 1,828 unigenes was shown in differential expression by comparison of SK *vs.* TK, SK *vs.* ZK, SK *vs.* LK, TK *vs.* ZK, TK *vs.* LK and ZK *vs.* LK. We envision that these transcriptome datasets could serve as an important public information platform to accelerate further studies of *P. ternate*. Moreover, further analysis of candidate genes encoding enzymes that potentially involved in *P. ternate* total alkaloids, acid and other active ingredients biosynthesis could be rapidly identified and validated using this approach. The transcriptome sequence of *P. ternate* provides a rich resource of genetic information for synthesis and regulation of bioactive substances, breeding and further genetic improvement of Araceae family.

Abbreviations

NGS, Next-Generation Sequencing; GAP, good agricultural practices; FPKM, fragments per kilobase of exon mode per million mapped reads; NCBI, National Center for Biotechnology Information; ORFs, open reading frames; GO, gene ontology; COG, Clusters of Orthologous Group; KEGG, Kyoto Encyclopedia of Genes and Genomes; KO, KEGG Orthology; SSRs, sequence repeats; SNP, Single Nucleotide Polymorphism

Declarations

Acknowledgements

Not applicable.

Authors’ Contributions

YJ and ML designed the study. XH and ZW conducted the literature search. YJ, JL and YL drafted the manuscript and prepared tables and figures. QL, SW, YW and ZW contributed to revisions of the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by the National Standardization of Traditional Chinese Medicine Project (ZYBZH-Y-SC-40), Scientific research promotion plan of Xinglin Scholars (CXTD2018016) and National Natural Science Foundation of China (81173492).

Availability of data and materials

The dataset supporting the conclusions of this article is included within the article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing of interests

The authors declare that they have no competing interests.

Author details

[1] College of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China. [2] PU-UM Innovative Institute of Chinese Medical Sciences, Guangdong-Macau Traditional Chinese Medicine Technology Industrial Park Development Co., Ltd, Hengqin New Area, Zhuhai, China.[3] State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao, China.

References

1. Wu X, Wang SP, Lu JR *et al*. Seeing the unseen of Chinese herbal medicine processing (Paozhi): advances in new perspectives. *Chin Med-Uk* 2018, 13.
2. Xu JY, Dai C, Shan JJ *et al*. Determination of the effect of Pinellia ternata (Thunb.) Breit. on nervous system development by proteomics. *J Ethnopharmacol* 2018, 213:221-29.
3. Zhang ZH, Zhao YY, Cheng XL *et al*. General toxicity of Pinellia ternata (Thunb.) Berit. in rat: A metabonomic method for profiling of serum metabolic changes. *J Ethnopharmacol* 2013, 149(1):303-10.
4. Chen FJ, Zhang LX, Zhao CM. Isolation and characterization of microsatellite markers for Pinellia ternata and cross-species amplification. *Mol Ecol Resour* 2008, 8(6):1460-62.
5. Chung HS, Um JY, Kim MS *et al*. Determination of the site of origin of Pinellia ternata roots based on RAPD analysis and PCR-RFLP. *Hereditas* 2002, 136(2):126-29.
6. Xu T, Wang B, Liu XF *et al*. Microarray-based identification of conserved microRNAs from Pinellia ternata. *Gene* 2012, 493(2):267-72.
7. Margulies M, Egholm M, Altman WE *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437(7057):376-80.
8. Huse SM, Huber JA, Morrison HG *et al*. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007, 8(7).
9. Novaes E, Drost DR, Farmerie WG *et al*. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *Bmc Genomics* 2008, 9.
10. Grabherr MG, Haas BJ, Yassour M *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011, 29(7):644-52.
11. Mortazavi A, Williams BA, Mccue K *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008, 5(7):621-28.
12. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. *Trends Genet* 2000, 16(6):276-77.

13. Conesa A, Gotz S, Garcia-Gomez JM *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21(18):3674-76.

14. Conesa A, Gotz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 2008, 2008:619832.

15. Myhre S, Tveit H, Mollestad T *et al*. Additional Gene Ontology structure for improved biological reasoning. *Bioinformatics* 2006, 22(16):2020-27.

16. Ye J, Fang L, Zheng HK *et al*. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 2006, 34:W293-W97.

17. Tatusov RL, Galperin MY, Natale DA *et al*. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28(1):33-36.

18. Moriya Y, Itoh M, Okuda S *et al*. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, 35:W182-W85.

19. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol* 1999:138-48.

20. Temnykh S, DeClerck G, Lukashova A *et al*. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* 2001, 11(8):1441-52.

21. Yu JK, La Rota M, Kantety RV *et al*. EST derived SSR markers for comparative mapping in wheat and rice. *Mol Genet Genomics* 2004, 271(6):742-51.

22. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 2005, 23(1):48-55.

23. Varshney RK, Sigmund R, Borner A *et al*. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci* 2005, 168(1):195-202.

24. Iyer MK, Chinnaiyan AM. RNA-Seq unleashed. *Nat Biotechnol* 2011, 29(7):599-600.

25. Kanehisa M, Goto S, Sato Y *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012, 40(D1):D109-D14.

26. Xu YB, Crouch JH. Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci* 2008, 48(2):391-407.

27. Tanksley SD. Molecular markers in plant breeding. *Plant Molecular Biology Reporter* 1983, 1(1):308.

28. Thiel T, Michalek W, Varshney RK *et al*. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* 2003, 106(3):411-22.

29. Untergasser A, Cutcutache I, Koressaar T *et al*. Primer3-new capabilities and interfaces. *Nucleic Acids Res* 2012, 40(15).

30. Kim D, Landmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015, 12(4):357-U121.

31. McKenna A, Hanna M, Banks E *et al*. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010, 20(9):1297-303.

32. Tarazona S, Garcia-Alcalde F, Dopazo J *et al*. Differential expression in RNA-seq: A matter of depth. *Genome Res* 2011, 21(12):2213-23.

33. Audic S, Claverie JM. The significance of digital gene expression profiles. *Genome Res* 1997, 7(10):986-95.

34. Huang X, Jing Y, Liu DJ *et al*. Whole-transcriptome sequencing of Pinellia ternata using the Illumina platform. *Genet Mol Res* 2016, 15(2).

35. Tu K, Yu H, Guo Z *et al*. Learnability-based further prediction of gene functions in Gene Ontology. *Genomics* 2004, 84(6):922-8.

36. Gene Ontology C. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 2004, 32(90001):258D-61.

37. Koudounas K, Manioudaki ME, Kourti A *et al*. Transcriptional profiling unravels potential metabolic activities of the olive leaf non-glandular trichome. *Front Plant Sci* 2015, 6:633.

38. Xu Y, Zeng X, Wu J *et al*. iTRAQ-Based Quantitative Proteome Revealed Metabolic Changes in Winter Turnip Rape (Brassica rapa L.) under Cold Stress. *Int J Mol Sci* 2018, 19(11).

39. Liang Z, Zhang J, Wong L *et al*. Characterization of Secondary Metabolites from the Raphides of Calcium Oxalate Contained in Three Araceae Family Plants Using Laser Microdissection and Ultra-High Performance Liquid Chromatography-Quadrupole/Time of Flight-Mass Spectrometry. *European Journal of Mass Spectrometry* 2013, 19(3):195-210.

Supplemental Information Note

The additional file mentioned on page 4 was omitted by the authors in this version of the paper.

Figures

Figure 1

Medicinal materials of Pinelliae Rhizoma.

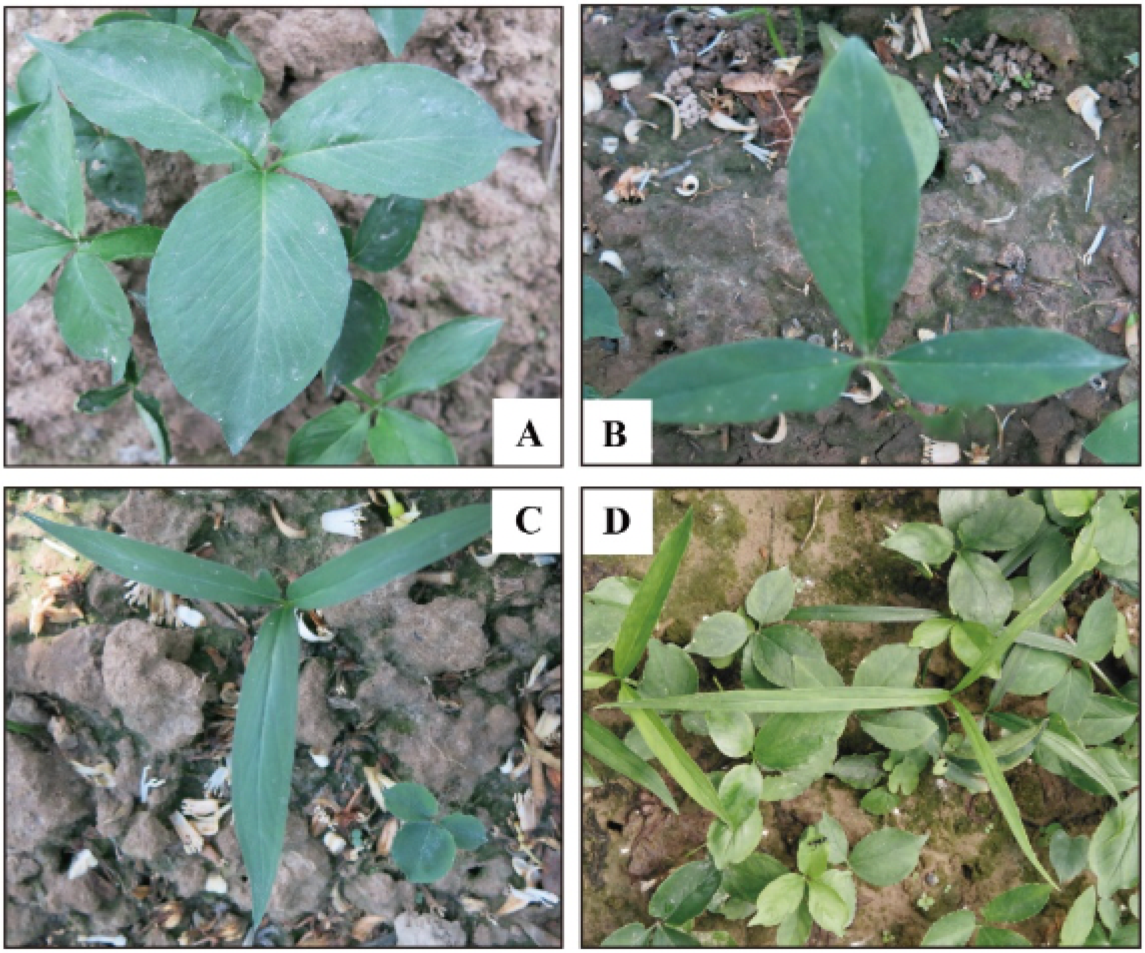

Figure 3

Different leaf shapes of four genotypes of P. Ternate. (A) SK, (B) TK, (C) ZK and (D) LK.

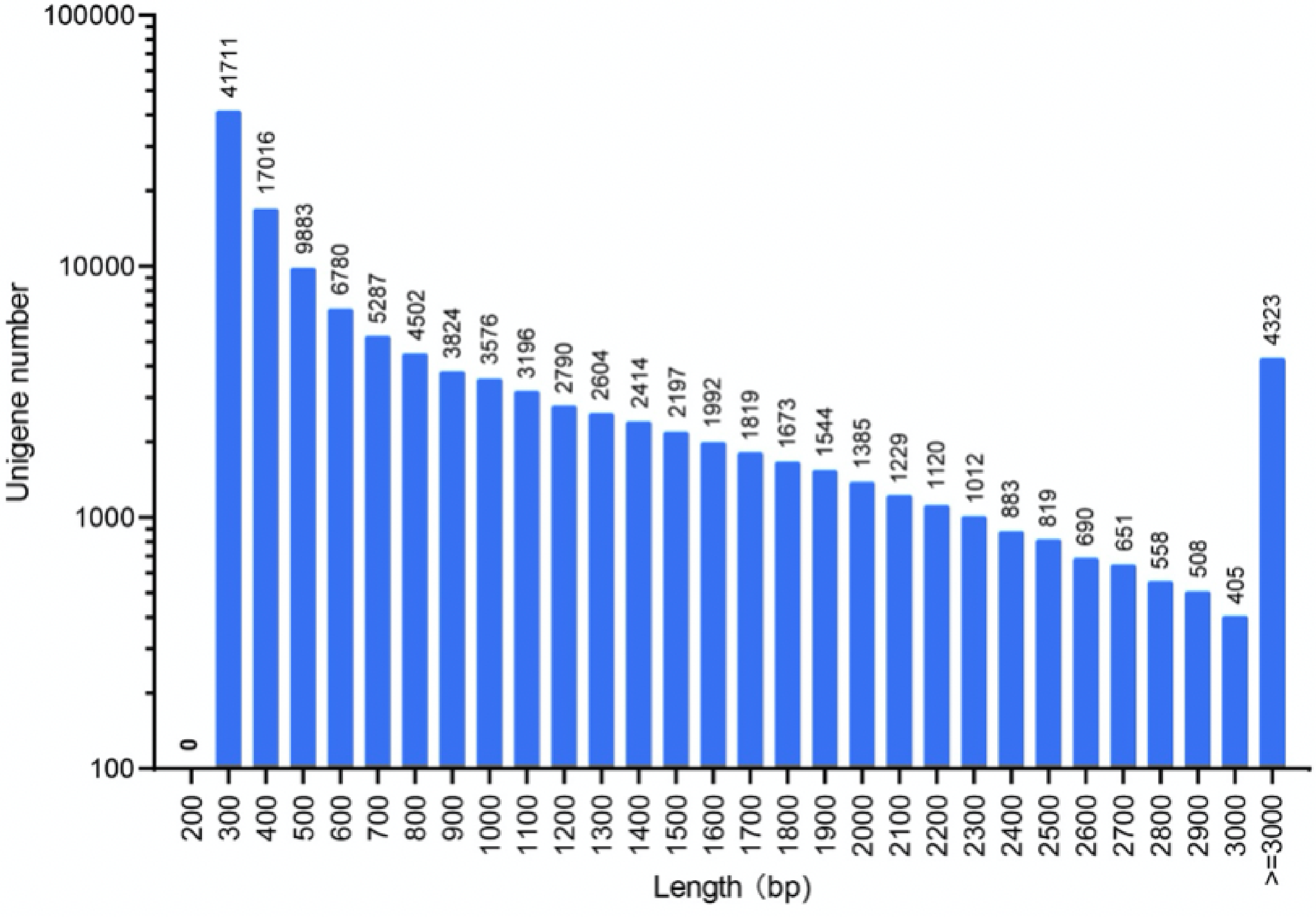

Figure 5

Length distribution of P. Ternate unigenes.

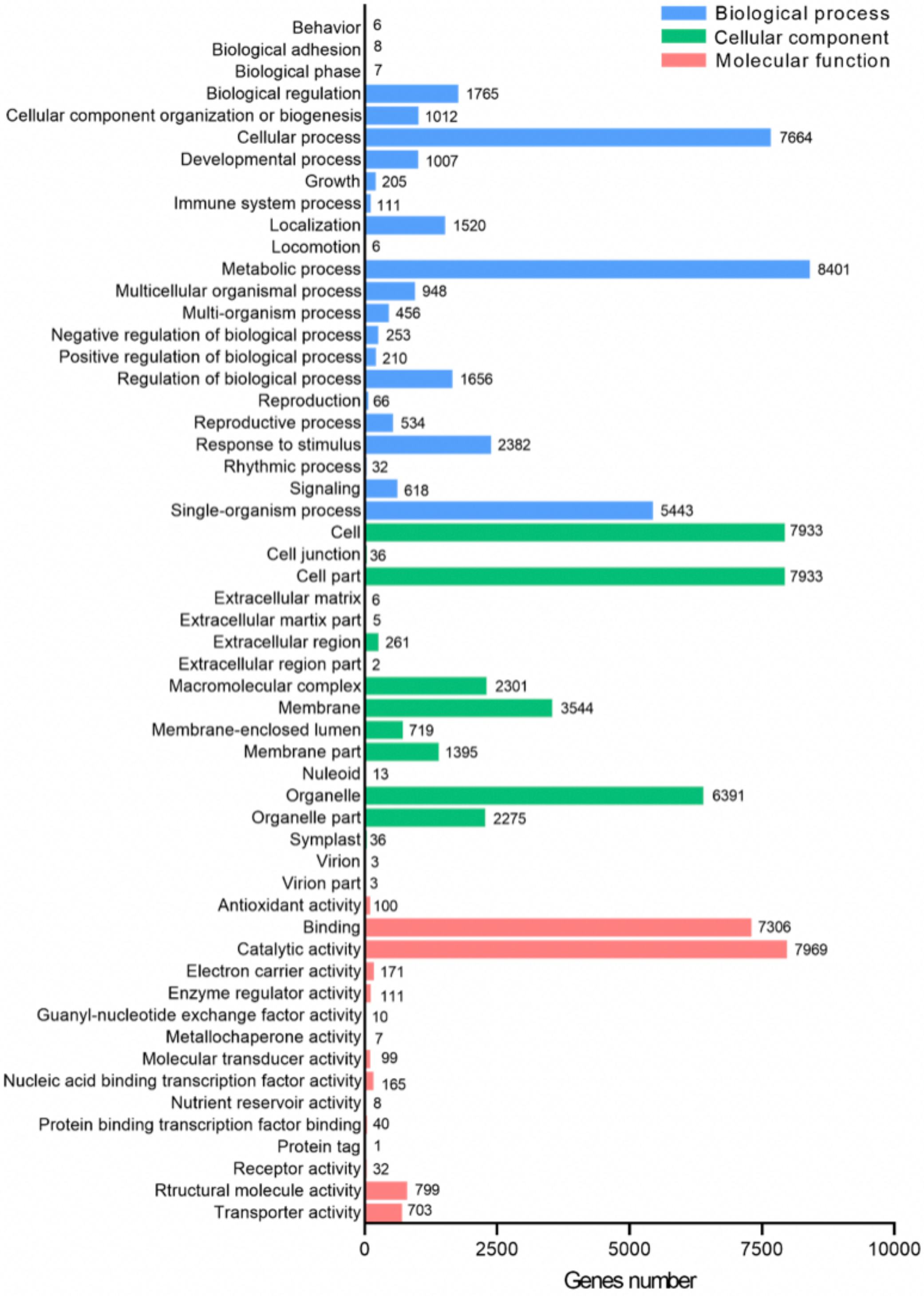

Figure 7

Histogram of gene ontology classification. The results are summarized in three main categories, including biological process, cellular component and molecular function.

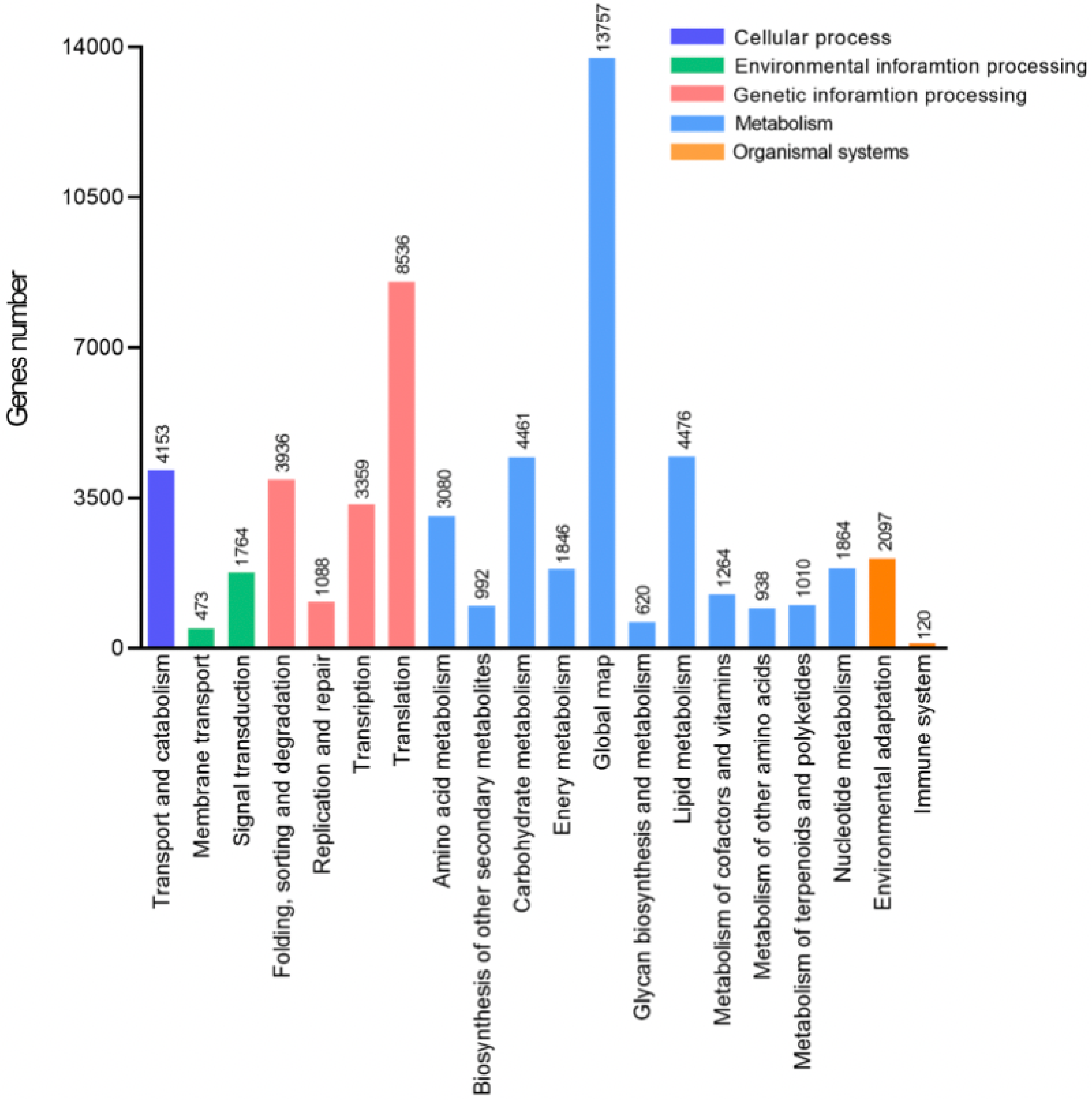

Figure 9

KEGG biochemical maps of P. Ternate unigenes.