Full-Length Transcriptome Sequencing and EST-SSR marker development of tiger lily (*Lilium lancifolium* Thunb.)

Mingwei Sun (sunmingweihappy@163.com)
Lianyungang Academy of Agricultural Sciences
Yilian Zhao
Nanjing Agricultural University
Xiaobin Shao
Lianyungang academy of agricultural science
Jintao Ge
Lianyungang academy of agricultural science
Xueyan Tang
Lianyungang academy of agricultural science
Pengbo Zhu
Lianyungang academy of agricultural science
Jiangying Wang
Lianyungang academy of agricultural science
Tongli Zhao
Lianyungang Academy of Agricultural Sciences

Research Article

Keywords: tiger lily, transcriptional diversity, transcriptome information

Posted Date: April 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-415397/v1

Abstract

It is well known that transcriptional diversity plays important roles in plant biological regulation. But for the difficulty in full-length transcripts obtainment, the available tiger lily (*Lilium lancifolium Thunb*) transcriptome characterization are still not complete. To improve the integrity of tiger lily transcriptome information, (SMRT PacBio single-molecule long-read sequencing technology) was employed to accomplish the whole transcriptome profiling. A total of 815,624 CCS (Circular Consensus Sequence) reads with mean length of 1,295 bp were obtained. Based on these transcripts, 61,744 reads were full-length reads containing both the 5’ primer, 3’ primer and the poly (A) tail and 3,319 EST-derived SSRs were developed from 2968 unigenes. With the obtained informative reference transcriptome,768 transcription factors and 6,852 long non-coding RNAs were identified, providing a comprehensive framework of the transcriptional regulation network. Of all the annotated transcripts, 15,608 were distributed into 25 various Clusters of euKaryotic Orthologous Groups (KOG), and 10,706 unigenes were categorized into 52 functional groups which were divided into three categories. These results would provide a comprehensive set of reference transcripts and further improve our understanding of the tiger lily transcriptomes.

Introduction

Owning different corolla shape and color, tiger lily was world widely welcome for its beautiful appearance. On the other hand, tiger lily is also considered as the healthy food and taken as one of the medical plants published by the Chinese Ministry of Health for the reason that it contains various of nutrients and bioactive compounds such as phenolic glycosides, pectin, steroidal saponins, alkaloids, and kinds of vitamins. As a triploid plant, tiger lily owns an enormous genome about 32.8 pg – 47.9 pg (Leitch et al. 2007). Information of the whole genome and transcriptome would make great contribution for tiger lily gene model study, but the large and not unrevealed genome and high heterozygosity make it difficult for gene functional study.

Next-generation sequencing platforms have provided us active transcriptional patterns in transcriptomic analysis over the past decade. Despite gene expression could be accurately quantified with the massive resequencing throughput, some important transcript information covering long lengths, including AS (alternative splicing), SSR (ssimple sequence repeats), and lncRNA (long non-coding RNA), would be lost for the reason that RNA or cDNA need to be fragmented during sample preparation and only information with short-reads could be obtained with the next-generation sequencing platforms. With the research requirement, the long-read sequencing platforms that could provide entire cDNA molecules are available, including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) (Rhoads, Au 2015; Bayega et al. 2018). Because the full length of transcripts could be captured by PacBio, the accuracy of gene annotation, isoform identification, and lncRNA discovery was improved much compared with those from the next-generation sequencing platforms (Abdel-Ghany et al. 2016; Gonzalez-Garay 2016). Thus, PacBio single-molecule real-time isoform sequencing has been widely used for transcriptome profiling in many plants and animals (Dóra et al. 2018; Dahe et al. 2019).

In this study, the PacBio SMRT technology was adopted to carried out the full-length cDNA library sequencing of tiger lily. To obtain a transcriptome with high accuracy, we combined the RNA-seq from Illumina HiSeq plat to achieve comprehensive full-length transcriptomes for isoform transcripts identification and quantification. This dataset could provide us rich information about full-length cDNA sequences that can help broaden our understanding of tiger lily. The function annotation of these full-length transcripts and detection of lncRNAs (Long noncoding RNAs) would help us to further understand the gene functions. The developed SSR primers in this study would be useful tools in gene mapping, affinity and systematic classification, and make contribution for tiger lily breeding. Therefore, These results would provide efficient tools for gene study and breeding.

Materials And Methods

Plant material, RNA was extraction, cDNA library preparation, and PacBio ISO-Seq

Tiger lily (*Lilium lancifolium* Thunb.) was grown at Lianyungang Academy of Agricultural Sciences (Lianyungang, China). Samples of the scales from three individual plants were collected and frozen in liquid nitrogen and stored at −70 °C for RNA extraction. The Iso-Seq library was constructed using the Isoform Sequencing protocol (Iso-Seq) with the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as described by Pacific Biosciences (PN 100-092-800-03).

PacBio ISO-Seq analysis

Circular consensus sequence (CCS) was generated after the raw reads were processed with the SMRTlink 7.0 software (https://www.pacb.com/support/software-downloads/). Then the CCSs were classified into full length and non-full length reads according to the rule that whether the 5'primer, 3'primer and poly A tail were existed. Clustering was completed to predict consensus isoforms from full-length transcripts using the Interactive Clustering and Error Correction (ICE) algorithm and the consensus isoforms polishing with non-full-length transcripts was carried out with software Quiver v1.1.0 (Pacific Biosciences of California Inc.; Menlo Park, CA, USA) (Chin et al. 2013). The obtained sequences without redundancy and extention on either end were defined as transcripts. To ensure the sequence accuracy, PacBio reads were corrected using the Illumina RNAseq data with the software LoRDEC (Leena et al. 2014).After all redundancy in corrected consensus reads was removed by CD-HIT (Li, Godzik 2006), the final consensus isoforms for the subsequent analysis were obtained.

Annotation of the transcriptome

The isoforms were subjected to databases of the NR (NCBI non-redundant protein sequences) (Pruitt, Tatusova, Maglott 2005), COG (Clusters of Orthologous Groups of proteins) (Tatusov et al. 2003), KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa 2002), and Swiss-Prot (A manually annotated and reviewed protein sequence database) (Boeckmann et al. 2003) with software diamond (v0.8.36) (E-value $\leq 1\times10^{-5}$) (Benjamin et al. 2015) and to NT database (NCBI nucleotide sequences) with program

ncbi-blast (v2.7.1). Blast2GO (v2.5.0) (Conesa et al. 2005) and HMMER (v3.1) (Eddy 1998) were used to analyze GO term annotations and Pfam database (Finn et al. 2014), respectively.

Detection of CDS, TF, lncRNA, and SSR

Softwares of ANGLE (Shimizu, Adachi, Muraoka 2006), iTAK (Zheng et al. 2016), MISA (Beier et al. 2017) were employed to carried out the prediction for CDS (Coding DNA Sequence), TF (Transcription factor), and Simple SSR (Sequence Repeat). The lncRNAs were obtained after all the transcripts with coding potential were screened by CNCI (Coding– Non-Coding Index) (Sun et al. 2013), CPC2 (Coding Potential Calculator 2) (Kang et al. 2017), Pfam (Finn et al. 2014), and PLEK (https://sourceforge.net/projects/plek/) sequentially. Software MISA (V1.0) was employed for SSRs detection (Sebastian et al. 2017). The parameters were set for identifcation of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide unit size with a minimum repeats of of 10, 6, 5, 5, 5, and 5, respectively. SSR primers were designed by Batch Primer 3 (You et al. 2008). PCR amplifications with the tiger lily DNA template and PCR products separation with 8% polyacrylamide gels were carried out to test the SSR primers usability.

Results

Full-length transcripts from the sepals of *Lilium lancifolium Thunb*

To obtain a comprehensive full-length transcriptome data, the extracted RNA from five samples were pooled together and sequenced using the PacBio Sequel platform. A total of 54.86 G subread bases was obtained by two SMRT cells from the constructed PacBio library, yielding 815,624 CCS (Circular Consensus Sequence) reads with mean length of 1,295 bp (Table 1). Of the total reads, 461,744 (56.62%) were full-length reads containing the 5’ primer, 3’ primer and the poly (A) tail and 459,322 (56.32%) were FLNC (full length readsnon-chimeric) sequences (Table 1). With ICE Quiver and Arrow polishing algorithms, 38,707 polished full-length consensus transcripts with a mean length of 1574 bp were produced (Supplementary material 1). To improve consensus accuracy, FLNC sequences were corrected with Illumina short-read RNA-seq reads with software LoRDEC. As a result, 8,596 nucleotides in the 38,707 transcripts were revised, and the mean length is still 1574 bp (Table 1). We found that suquence length of most transcripts (98.3%) ranges from 0.2 to 4 kb (Supplementary material 1). With software CD-HIT, the redundant and similar sequences for one unigene was removed. Finally, a total of 15,608 unigenes were obtained with the threshold of 95% of software CD-HIT, and mean length of the unigenes turns to be 1627 (Supplementary material 2).

Table 1 Summary of reads of inserts from PacBio single-molecule long-read sequencing.

species	Subreads base (G)	Reads of CCS	nonfull-length reads	Full-length reads	FLNC reads	Consensus transcripts	Mean length (bp)	Unigenes
Lilium davidii var. unicolor Salisb	54.86	815624	353880	461744	459322	38707	1574	38707

Transcription factor prediction

As the key regulators for gene expression through activating or inhibiting gene transcription, transcription factors play important roles in plant growth, development, and reactions against various biotic and abiotic stresses. In this study, 768 transcription factors were identified and classified into 30 families (Supplementary material 3). The TFs in the tiger lily transcriptome mainly belonged to the MADS-MIKC (64, 8.33%), MADS-M-type (47, 56.12%), MYB-related (44, 5.73%), C3H (42, 5.47%), FAR1 (42, 5.47%), bZIP (42, 5.47%), C2H2 (41, 5.34%), GRAS (40, 5.21%), bHLH (40, 5.21%), AP2/ERF-ERF (37, 4.82%), and MYB (32, 4.17%) families (Fig. 1) . Gene sequence character of the TFs would help us much in gene functional studie and serve as an excellent starting point for rapid gene discover.

Simple sequence repeats (SSRs) analysis

With the advantages of codominance, stability, high reproducibility, high polymorphism, primers specificity, and easy detection, SSR markers play a vital role in high-density genetic map construction, DNA fingerprinting building, germ-line purity and phylogenetic evolution analysis. In this research, 3319 SSRs were detected from 2968 unigenes. Among these transcripts, 2675 transcripts only contains one single SSR and 293 transcripts contains 2 or more SSR loci. Numbers of mono-, di-, tri-, repeats were 1039, 862, and 766 respectively, accounting for the most frequent motif type (Fig. 2, Supplementary material 4). According to the SSR primer information, 128 pairs of primers were developed and all these primers could successfully amplify the cDNA but only 105 (82.03%) primer pairs could successfully amplified PCR products when the whole genome DNA was taken as template.

Long non-coding RNA identifcation

LncRNA are a kind of poly-A noncoding RNAs that determine gene expression through affecting dosage compensationeffect, epigenetic regulation, cell cycle regulation and cellular differentiation regulation. To get a comprehensive LncRNA data, four computational approaches were used to identify lncRNAs, involving the CNCI (Coding-Non-Coding Index), Pfam (Protein family), PLEK (predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme), and CPC (Coding Potential Calculator) databases. With screening by the above four database, a total 824, 3147, 5457, and 1962 lncRNAs were identified from database CNCI, Pfam, PLEK, and CPC, respectively (Supplementary material 5). Of all the 6852 lncRNAs, 570 were considered commonly shared by the four database.

Functional annotation of transcripts

Sequence of the unigenes were functionally analyzed and classifed using NR (NCBI non-redundant protein sequences) , SwissProt, KEGG (Kyoto Encyclopedia of Genes and Genomes), KOG (euKaryotic Ortholog Groups), GO (Gene Ontology), NT (NCBI nucleotide sequences), and Pfam (Protein family) database. Among the 15608 unigenes identified 14614, 12575, 14554, 9650, 10705, 11309, and 10705 were annotated in the NR, SwissProt, KEGG, KOG, GO, NT, and Pfam database, respectively (Fig. 3). And 6656 unigenes were common shared by the above 7 databases. Species distribution annotation

illustrates that the largest proportion of unigene group is Elaeis guineensis (32.65%), followed by Phoenix dactylifera (25.82%), Musa acuminata (8.11%), Ananas comosus (7.15%), Asparagus officinalis (5.29%), Nelumbo nucifera (2.30%), Anthurium amnicola (2.20%), Vitis vinifera (0.79%), and Lilium longiflorum (0.72%) (Fig. 4).

GO classification

GO analysis about the unigenes was carried out to further classify the gene functions. GO enrichment illustrated that 10,706 unigenes were categorized into 52 functional groups, which could be divided into three categories: biological process, cellular component and molecular function. In the biological process group, metabolic process accounts for the most unigenes (5188), followed by cellular process (4676) and single-organism process (3044). The largest processes in molecular function group are binding (6393), catalytic activity (4851), and transporter activity (473) in turn. Cell and cell part owns the same number (2258) of Cellular Component processes, followed by organelle (1594), macromolecular complex (1282), and membrane (1136) (Fig. 5).

KEGG classification

A total of 14554 unigenes were mapped into 354 KEGG functional pathway (Supplementary material 6), and allthe pathways were then divided into the following 6 groups: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, and Organismal Systems. Among these pathways, Ribosome (434) and Carbon metabolism (345) are the most dominant pathways, followed by Biosynthesis of amino acids (307), Protein processing in endoplasmic reticulum (253), and Spliceosome (236) (Supplementary material 7). The KEGG functional classification provided valuable informations for investigating particular processes, functions, and pathways in tiger lily .

KOG classification

To further research the functional annotation and classification of the tiger lily unigenes, all unigenes were analyzed with the Clusters of KOG database. Result shows that 9650 unigenes could be classified into 25 categories. Of all categories, General function prediction accounting for the most unigenes (1709, 15.88%), followed by Posttranslational modification, protein turnover, chaperones (1303, 12.11%), Signal transduction mechanisms (843, 7.83%), and Translation, ribosomal structure and biogenesis (771, 7.16%). Percentages of five groups were less than 1.00%, including Nucleotide transport and metabolism, Defense mechanisms, Nuclear structure, Extracellular structures, and Cell motility (Supplementary material 8).

Discussion

As an efficient and reliable method for full-length transcripts screening, SMRT sequencing technology has made great contribution for whole-transcriptome profiling studies, especially for some plant species without reference genome sequences. In this study, the tiger lily whole-transcriptome equencing was

carried out with SMRT sequencing technology based on the PacBio Sequel platform. To ensure the sequencing accuracy, a total of 63.85 Gb sequencing data was generated, including 815,624 CCS and 459,322 FLNC reads. Percentage of FLNC reads in all CCS reads was 56.31%, which is similar with the result that obtained in alfalfa (Chao et al. 2019) and strawberry (Li et al. 2017). Although gene expression could be accurately quantified with next-generation sequencing platforms, the long nucleotide sequences usually could not be captured. Transcripts obtained by SMRT sequencing technology are commonly longer than that from next-generation sequencing platforms, where one read usually represents a full-length transcript (Sharon et al. 2013). In this study, the average length of tiger lily transcripts by SMRT was 72,076 bp, which is much longer than those obtained by Illumina sequencing technology from previous research (114.59 bp – 297.76 bp) (Li et al. 2014; Tian, Jihua, Xie 2019). Furthermore, we found that 97.45% of all transcripts were longer than 2,000 bp in this SMRT sequencing result, indicating that PacBio SMRT sequencing technology is an efficient and appropriate approach for transcript sequences information study, especially for the long transcript sequences.

Transcription factors are a kind of important regulatory factors that affecting gene expression. With lncRNA, microRNA, and methylation together, transcription factors play a regulatory role in fine-tuning gene expression in cell differentiation and growing development, especially in response to abiotic stress and plant disease (Feng et al. 2020; Gong et al. 2020) (Arnold et al. 2020). An increasing number of studies were focused on regulatory network in plants in the recent research. However, no lncRNAs from tiger lily have been reported until now. In this study, 768 transcription factors and 6852 lncRNAs were identified from the tiger lily transcripts, which will be useful for further research in tiger lily.

For the reason the tiger lily genome is not avaliable now, the SSR primers would be useful tools in gene mapping, genetic diversity, comparative genomics, and gene functional study, especially for the breeding work. The long transcript sequence information made it easier for high-quality SSR maker development. Based on the polished gene sequence, 3319 SSRs were detected from 2968 SSR-containing unigenes. Among the six types of SSR repeat motifs (motif with 1 to 6 nucleotides), mononucleotide repeats account for the biggest proportion, accounting for a proportion of about 41.32%, followed by di- and tri-nucleotide motifs. In the present study, the most frequent mono-, di-, and tri-nucleotide motifs were A/T, AT/AT and CCG/CGG, respectively. Of all the nucleotide repeats, the most abundant repeat motif is A/T, which accounting about for 39.59% of all the motifs and is thought to be frequent in the genomic sequences of plants (Ulf, Hans, Leif 1993), followed by AT/AT (15.94%) and AG/CT (14.41%) belonging to the di-nucleotide motifs. Of the tri-nucleotide motifs, CCG/CGG accounts for a proportion of 9.01% of the total motifs, followed by AGG/CCT (3.93%) and AGC/CTG (3.65%). According to the SSR marker information, 128 pairs of primers were developed to test the maker reliability. As a result, when the cDNA, which obtained through reverse transcription of RNA, was taken as PCR amplification temple, all the 128 pairs of primers could successfully amplified PCR products. But when the whole genome DNA was taken as template, only 105 (82.03%) primer pairs could successfully amplified PCR products. The targeting region with long-length intron and primer nucleotide located on two neighboring exons may be the reasons for the above failed PCR amplification of 23 pairs of primers. These results suggested that the developed SSR markers based on PacBio SMRT sequencing platform is an effective and helpful tools for

genetic diversity, gene mapping, and other types of genetic studies in tiger lily, and would make great contribution in the tiger lily breeding work.

In conclusion, the full-length transcriptome of tiger lily based on the PacBio SMRT sequencing technology was analyzed. This study provides the first the third-generation long-read transcriptome sequencing of tiger lily. Based on the obtained transcriptome data, 38,707 unigenes, 6852 lncRNAs, and 768 TF members were identified. In addition, 3,319 SSRs were detected, and 105 from 128 primer pairs could successfully amplified PCR products. The obtained transcriptome data and developed SSR markers may facilitate further genetic studies and help to breeding new varieties on tiger lily.

Declarations

Authors' contributions

MS, YZ, and TZ designed the experiment and revised the manuscript. MS, XS, JG, and XT collected and prepared RNA samples for RNA-seq and performed some RT-PCR. PZ, JW, and TZ performed data analysis and wrote the manuscript. All authors read and approved the final manuscript.

References

Abdel-Ghany, SE, M Hamilton, JL Jacobi, P Ngam, N Devitt, F Schilkey, A Ben-Hur, ASN Reddy. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nature communications* 7:11706.

Arnold, A, EL Imada, ML Zhang, DP Edward, FJ Rodriguez. 2020. Correction to: Differential gene methylation and expression of HOX transcription factor family in orbitofacial neurofibroma. *Acta Neuropathologica Communications* 8.

Bayega, A, S Fahiminiya, S Oikonomopoulos, J Ragoussis. 2018. Current and Future Methods for mRNA Analysis: A Drive Toward Single Molecule Sequencing. *Gene Expression Analysis.*

Beier, S, T Thiel, T Munch, U Scholz, M Mascher. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33:2583-2585.

Benjamin, Buchfink, Chao, Xie, Daniel, Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 2015, 12(1): 59-60.

Boeckmann, B, A Bairoch, R Apweiler, M-C Blatter, A Estreicher, E Gasteiger, MJ Martin, K Michoud, C O'donovan, I Phan. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* 31:365-370.

Chao, Y, J Yuan, T Guo, L Xu, Z Mu, L Han. 2019. Analysis of transcripts and splice isoforms in Medicago sativa L. by single-molecule long-read sequencing. *Plant Molecular Biology*. 99(3): 219-235.

Chin, CS, DH Alexander, P Marks, AA Klammer, J Drake, C Heiner, A Clum, A Copeland, J Huddleston, EE Eichler. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* 10:563-569.

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

Dóra, T, B Zsolt, C Zsolt, S Michael, BI Zsolt. 2018. Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses. *Frontiers in Genetics* 9:259.

Dahe, Qiao, Chun, Yang, Juan, Chen, Yan, Guo, Li, Suzhen. 2019. Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (*Camellia sinensis*). *Scientific Reports*. 9(1): 1-13.

Eddy, SR. 1998. HMMER: profile HMMs for protein sequence analysis. *Bioinformatics*. 14. 755-763.

Feng, X, W Liu, H Dai, Y Qiu, G Zhang, ZH Chen, F Wu. 2020. HvHOX9, a Novel Homeobox Leucine Zipper Transcription Factor Revealed by Root miRNA and RNA Sequencing in Tibetan Wild Barley, Positively Regulates Al Tolerance. *Journal of Experimental Botany*.

Finn, RD, A Bateman, J Clements, P Coggill, RY Eberhardt, SR Eddy, A Heger, K Hetherington, L Holm, J Mistry. 2014. Pfam: the protein families database. Nucleic acids research 42:222-230.

Gong, J, H Fan, J Deng, Q Zhang. 2020. LncRNA HAND2-AS1 represses cervical cancer progression by interaction with transcription factor E2F4 at the promoter of C16orf74. J*ournal of Cellular and Molecular Medicine*.

Gonzalez-Garay, ML. 2016. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). *Transcriptomics and gene regulation. Springer, Dordrecht*, 2016: 141-160.

Kanehisa, M. 2002. The KEGG database. *Novartis Foundation Symposium* 247:91.

Kang, Y, D Yang, L Kong, M Hou, Y Meng, L Wei, G Gao. 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research* 45(W1): W12-W16.

Leena, Salmela, Eric, Rivals. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics*. 30(24): 3506-3514.

Leitch, IJ, JM Beaulieu, K Cheung, L Hanson, MA Lysak, MF Fay. 2007. Punctuated genome size evolution in Liliaceae. *Journal of Evolutionary Biology* 20:2296-2308.

Li, W, A Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.

Li, XY, CX Wang, JY Cheng, J Zhang, JATD Silva, XY Liu, X Duan, TL Li, HM Sun. 2014. Transcriptome analysis of carbohydrate metabolism during bulblet formation and development in Lilium davidii var. unicolor. *BMC Plant Biology*,14:358.

Li, Y, C Dai, C Hu, Z Liu, C Kang. 2017. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *Plant Journal* , 90(1): 164-176.

Pruitt, KD, T Tatusova, DR Maglott. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 33(suppl_1): D501-D504.

Rhoads, A, KF Au. 2015. PacBio Sequencing and Its Applications. *Genomics,Proteomics & Bioinformatics* 13:278-289.

Sebastian, B, T Thomas, M Thomas, S Uwe, M Martin. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33(16): 2583-2585.

Sharon, D, H Tilgner, F Grubert, M Snyder. 2013. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology* 31(11): 1009-1014.

Shimizu, K, J Adachi, Y Muraoka. 2006. ANGLE: A SEQUENCING ERRORS RESISTANT PROGRAM FOR PREDICTING PROTEIN CODING REGIONS IN UNFINISHED cDNA. *Journal of Bioinformatics & Computational Biology* 04:649-664.

Sun, L, H Luo, D Bu, G Zhao, K Yu, C Zhang, Y Liu, R Chen, Y Zhao. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research* 41(17): e166-e166.

Tatusov, RL, ND Fedorova, JD Jackson, AR Jacobs, B Kiryutin, EV Koonin, DM Krylov, R Mazumder, SL Mekhedov, AN Nikolskaya. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41-41.

Tian, X, YU Jihua, J Xie. 2019. Study on Key Anti-freeze Genes and Pathways of Lanzhou Lily(Lilium davidii var. unicolor) by Transcriptome Sequencing. *Guangdong Agricultural Sciences* 46(8): 35-43.

Ulf, L, E Hans, A Leif. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nuclc Acids Research* 21(5): 1111-1115.

You, FM, N Huo, YQ Gu, MC Luo, Y Ma, D Hane, GR Lazo, J Dvorak, OD Anderson. 2008. BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9(1): 253.

Zheng, Y, C Jiao, H Sun, HG Rosli, MA Pombo, P Zhang, M Banf. 2016. iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors,Transcriptional Regulators, and Protein Kinases. *Molecular Plant* 9(12): 1667-1670.

Figures

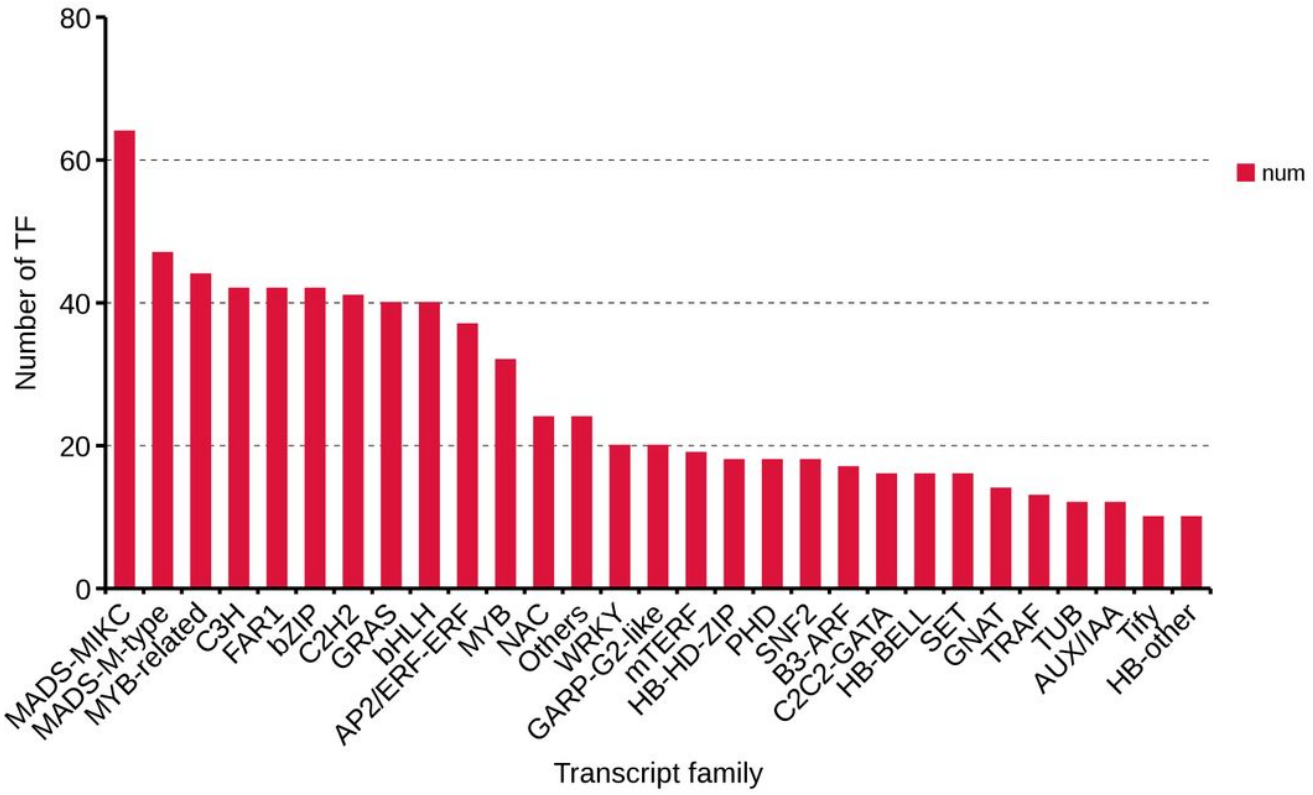

Figure 1

The TFs in the tiger lily transcriptome mainly belonged to the MADS-MIKC (64, 8.33%), MADS-M-type (47, 56.12%), MYB-related (44, 5.73%), C3H (42, 5.47%), FAR1 (42, 5.47%), bZIP (42, 5.47%), C2H2 (41, 5.34%), GRAS (40, 5.21%), bHLH (40, 5.21%), AP2/ERF-ERF (37, 4.82%), and MYB (32, 4.17%) families

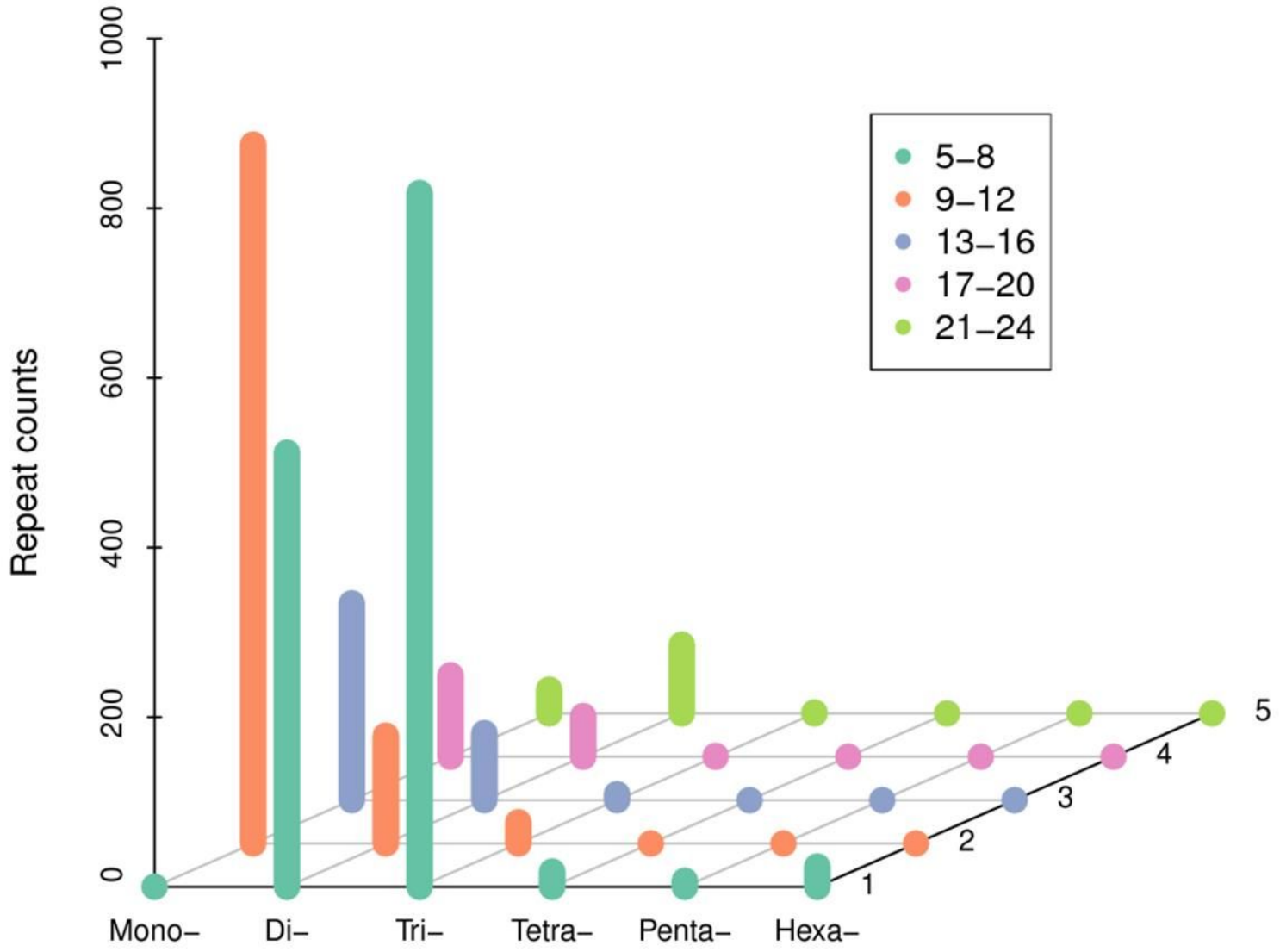

Figure 2

Numbers of mono-, di-, tri-, repeats were 1039, 862, and 766 respectively, accounting for the most frequent motif type

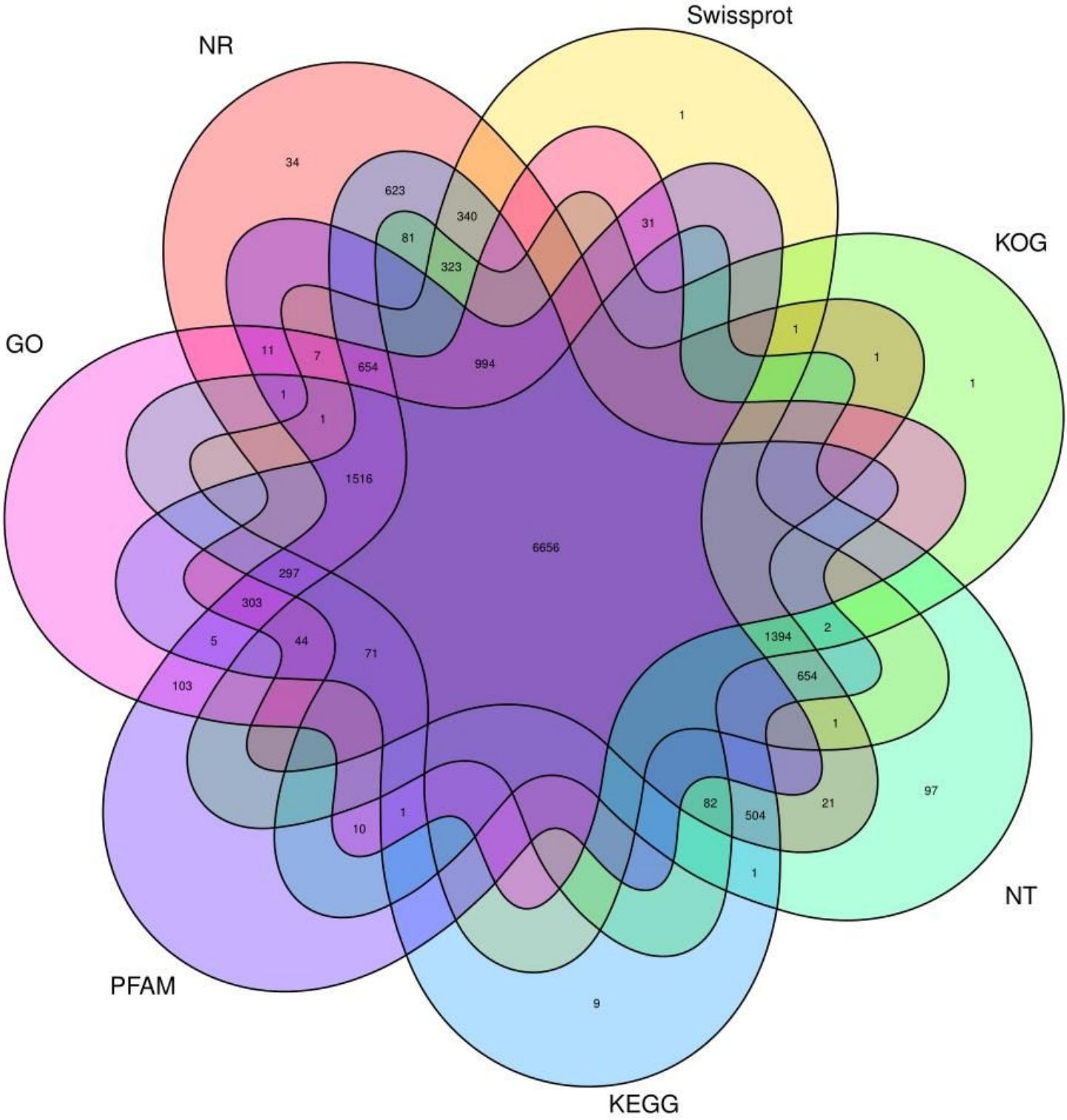

Figure 3

Among the 15608 unigenes identified 14614, 12575, 14554, 9650, 10705, 11309, and 10705 were annotated in the NR, SwissProt, KEGG, KOG, GO, NT, and Pfam database, respectively

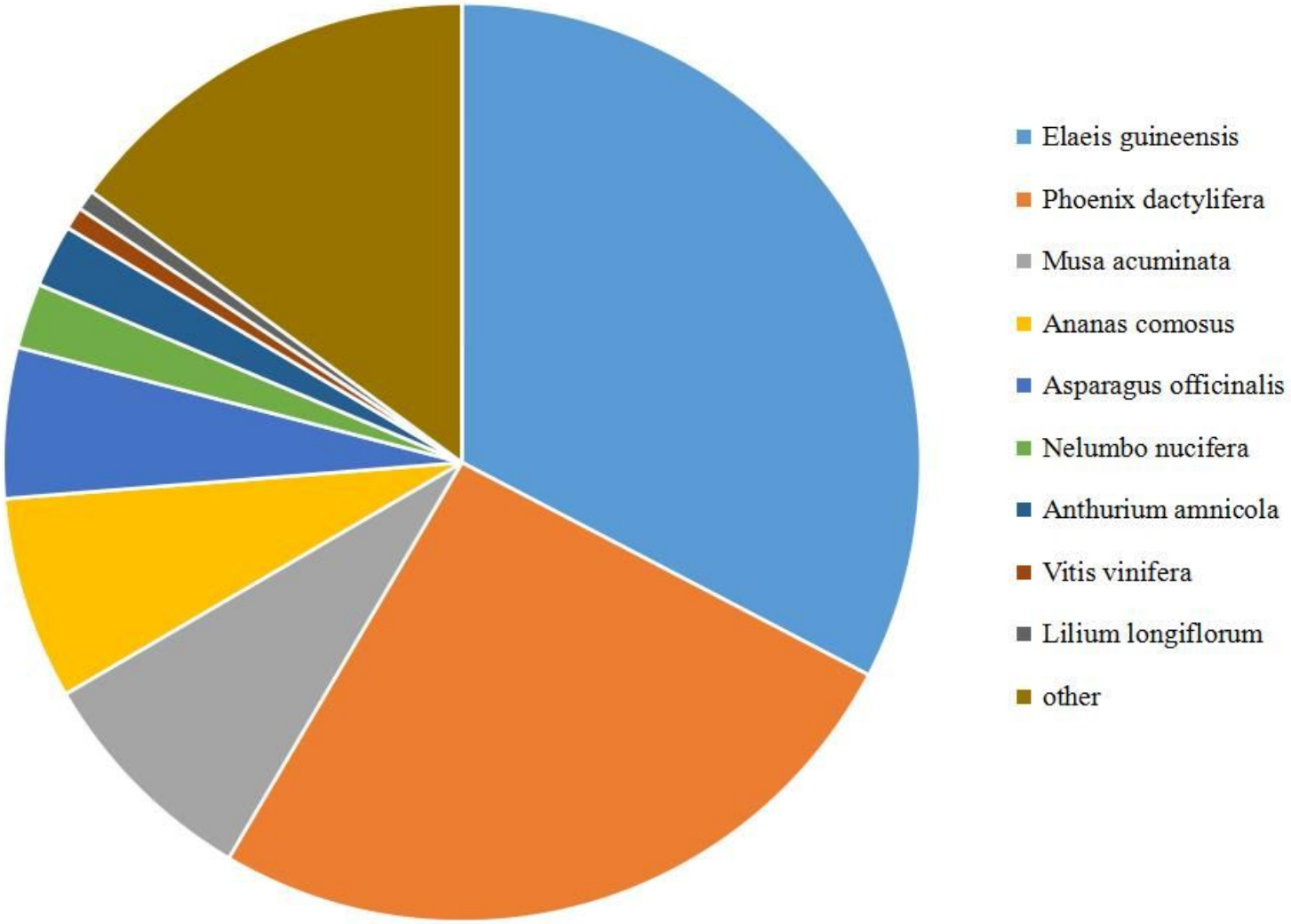

Figure 4

Species distribution annotation illustrates that the largest proportion of unigene group is Elaeis guineensis (32.65%), followed by Phoenix dactylifera (25.82%), Musa acuminata (8.11%), Ananas comosus (7.15%), Asparagus officinalis (5.29%), Nelumbo nucifera (2.30%), Anthurium amnicola (2.20%), Vitis vinifera (0.79%), and Lilium longiflorum (0.72%)

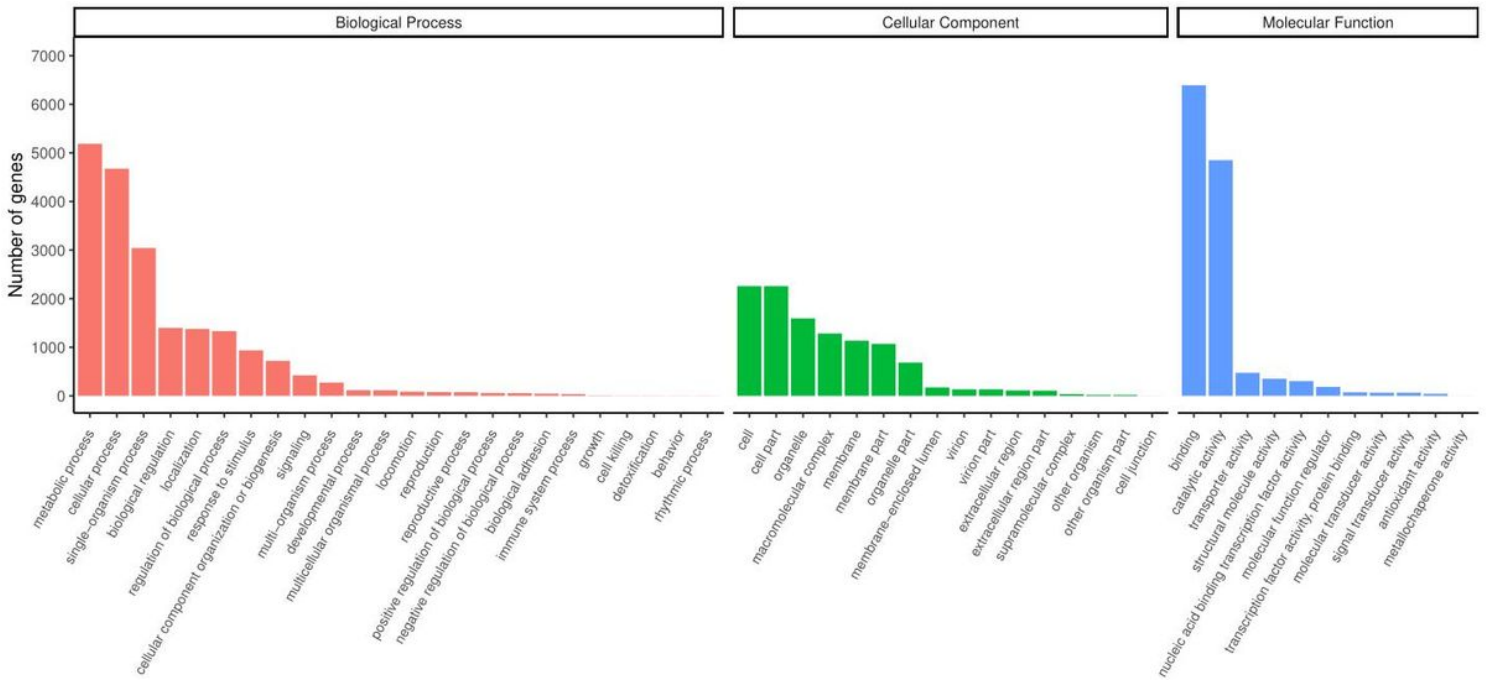

Figure 5

The largest processes in molecular function group are binding (6393), catalytic activity (4851), and transporter activity (473) in turn. Cell and cell part owns the same number (2258) of Cellular Component processes, followed by organelle (1594), macromolecular complex (1282), and membrane (1136)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial1.xls
- Supplementarymaterial2.xls
- Supplementarymaterial3.xls
- Supplementarymaterial4.xls
- Supplementarymaterial5.xls
- Supplementarymaterial6.xls
- Supplementarymaterial7.xls