De novo assembly of Chinese forest musk deer (*Moschus berezovskii*) transcriptome from Next-Generation mRNA Sequencing

Zhongxian Xu[1], Hang Jie[2], Binlong Chen[1], Uma Gaur[1], Mingyao Yang[1], Nan Wu[1], Jian Gao[1], Pinming Li[2], Guijun Zhao[2], Dejun Zeng[2] and Diyan Li[1]

[1]Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, China

[2] Laboratory of Medicinal Animal, Chongqing Institute of Medicinal Plant Cultivation, Nanchuan, Chongqing, China

Corresponding Author:

Diyan Li[1]

NO. 211 of Huimin Road, Wenjiang District, Chengdu, Sichuan Province, 611130, China

Email address: diyanli@sicau.edu.cn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

ABSTRACT

Musk secretion in male musk deer is regarded as a propitious mode of sexual election to attract a greater number of females. However, the genetic mechanisms of musk secretion are still poorly understood and unresolved making it necessary to elucidate the possible genetic mechanisms of musk formation. In the present study, we used heart and musk gland tissues from a male musk deer for next-generation mRNA sequencing, integrated with de novo assembly, unigenes annotation and differentially expressed genes analysis. A total of 239,383 transcripts and 208,730 unigenes were obtained from 2 pooled RNA samples. Annotated analysis indicated steroid compound metabolism (steroid biosynthesis, steroid hormone biosynthesis, aldosterone-regulated sodium reabsorption, terpenoid backbone biosynthesis) related to musk formation were annotated to many pathways; relevant genes were identified as well. In addition, 8,986 differentially expressed genes (6,068 up- and 2,198 down-regulated) between heart and musk gland were identified, among them, steroid component metabolism were abundant. Further exploration of functional enrichment analysis showed that pathways involved in musk secretion were up-regulated in musk gland compared with heart, especially steroid biosynthesis and terpenoid backbone biosynthesis whose metabolic productions were key components of musk. We identified several candidate genes such as *DHCR7*, *DHCR24*, *NSDHL*, *CYP3A5*, *FDFT1*, *FDPS* and *HMGCL* which were closely involved in metabolism of steroid, terpenoid and ketone bodies. Our data are expected to represent the most comprehensive sequence resource available for the forest musk deer so far, and provide a basis for further research on molecular genetics and functional genomics of musk secretion.

Keywords Chinese forest musk deer (*Moschus berezovskii*); Transcriptome; mRNA-seq; De novo assembly

INTRODUCTION

Chinese forest musk deer (*Moschus berezovskii*) is one of six species which are entirely Asian in their present distribution (excepting one is distinct in Europe where the earliest musk deer are known to have existed from Oligocene deposits) and famous for the secretion of musk. Musk is secreted by musk gland of sexually mature male individuals, it is used in Chinese traditional medicine and perfume manufacturing widely(Sheng 1996,, Su, Wang et al. 2001) because of its unique fragrance and significant role in anti-inflammation, anti-tumor, central nervous system and cardio-cerebral-vascular system(Cao and Zhou 2007, Feng and Liu 2015). Primary chemical ingredients of musk include: muscone, muscopyridine, cholesterol, cholesterol ester and male hormone (Li, Chen et al. 2016). Chinese forest musk deer was listed as endangered because of the serious decline in the population size caused by over-exploitation, shrinkage in distribution, habitat destruction and degradation. Therefore, it was cited as first class “key” species of wildlife protected by Chinese legislation in 2002(Guan, Zeng et al. 2009). Being a charismatic species Chinese forest musk deer has been the hot spot of research to study the ecology, domestication(Zhang, Deng et al. 1985, Deng 1986) and the pharmacological

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

functions of its musk (Seth, Mukhopadhyay et al. 1973, Feng and Liu 2015). The various aspects of the secretory mechanisms of musk have been explored by many research groups based on anatomy(Bi, Shen et al. 1984, Bi and Shen 1986), microsatellite(Guan, Zeng et al. 2009, Peng, Liu et al. 2009), mtDNA markers(Chen 2007, Zhao 2009), and microbiota analysis in recent years(Li, Chen et al. 2016). Unfortunately, the molecular and genetic mechanisms of the musk secretion have largely remained unexplored.

Transcriptome analysis is a preferred method to analyse functional genes in non-model organism(Vera, Wheat et al. 2008, Grabherr, Haas et al. 2011), which can overall identify and illustrate the mechanisms of important biological processes and incidences of diseases. Next-generation sequencing technology (RNA-Seq) has been a powerful tool for analysing the transcriptomes. This technology has been extensively used to study a number of species ranging from plants(Strickler, Aureliano et al. 2012), insects(Ya-Nan, Jun-Yan et al. 2013), birds(Luan, Liu et al. 2014), common marmoset(Maudhoo, Ren et al. 2014) and giant panda(Vera, Wheat et al. 2008) etc. In the absence of an appropriate reference genome/transcriptome, as is often the case in non-model organism studies, a de novo assembly is the only best option for sequence assembly (Strickler, Aureliano et al. 2012, Ockendon, O'Connell et al. 2015). Liu et al. reported the transcriptome expression in 5 developmental stages in Chinese sika deer antler(Yu, Baojin et al. 2013, Liu, Yao et al. 2014), which is a good example of sexual character in wild mammals. In the present study, a high throughput transcriptome sequencing (RNA-Seq) approach was adopted to investigate the transcriptome of heart and musk gland in mature male forest musk deer. Enormous gene fragments were obtained according to de novo assembly. Homology search with genes in public protein database and functional annotation such as NR, GO, KEGG were conducted on obtained unigenes. We also analysed the gene expression patterns in heart and musk gland that are involved in musk metabolism. These sequencing data and analysis provide a valuable resource for further research on the mechanisms of muck secretion.

MATERIALS AND METHODS

Sample collection and preparation

The heart and musk gland tissue were collected from one male forest musk deer after it died due to gastrointestinal illness in wild, from Chongqing Institute of Medicinal Plant Cultivation (Chongqing, China). The tissues were cut into small pieces and immediately stored in liquid nitrogen and then stored at -80℃ until RNA extraction.

RNA isolation, cDNA library construction and sequencing

We collected RNA from heart and musk gland. Total RNA was isolated from tissue samples using TRIzol reagent (Qiagen, Hilden, the Netherlands) following the manufacturer's instructions. The RNA quality and quantity were measured spectrophotometrically as well as by agarose gel electrophoresis. RNA samples having a RIN (RNA Integrity number) value of more than 8 were used for constructing 2 cDNA libraries. The samples for transcriptome analysis were prepared

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

using Illumina's kit following manufacturer's protocols and paired-end (PE) reads with 125bp read length of each were obtained by sequencing with the Illumina HiSeq 2500 platform. Sequence files corresponding to the heart and musk gland tissues from which the RNA originated, were generated in FASTQ format (sequence read and quality information in Phred format). The raw sequence data of Chinese forest musk deer have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra) with the accession number: PRJNA291827.

Read mapping and data processing

Quality checks and preprocessing were performed on raw sequence data with FASTQC software (http://www.bioinformatics.babraham. ac.uk/projects/fastqc/). Software cutadapt (https://pypi.python.org/pypi/cutadapt/1.2.1) was used for trimming adapters, and Prinseq（http://prinseq.sourceforge.net/）was used for quality control. The clean reads were aligned to NCBI database using blast+ (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download) for non-model organism, default parameters were used in all of softwares.

All the clean reads were pooled and subjected to transcriptome de novo assembly using Trinity (version trinityrnaseq_rr20140717) with the default settings (Grabherr, Haas et al. 2011). To further verify the sequencing accuracy and completeness, the BLASTn program was conducted to compare our sequences of Chinese forest musk deer with other species in EST database at NCBI. Gene expression levels were expressed as FPKM (Fragment Per Kilo bases per Million mapped Reads) (Mortazavi, Williams et al. 2008) using software RSEM(http://deweylab.biostat. wisc.edu/rsem/). BLASTX aligned with an E-value cut-off of 10^{-5} between unigenes and non-redundant protein sequences (NR) using Swissprot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) analysis was performed. Functional annotation with GO terms was conducted with Blast2go(Ana, Stefan et al. 2005), and then WEGO software(Jia, Lin et al. 2006) was used to do GO functional classification to understand the distribution of gene functions of Chinese forest musk deer. Annotation with KEGG pathway was performed by searching with BLASTX against the KEGG database.

Single nucleotide polymorphism(SNP) detection

In order to discover the genotype variation of different tissues on mRNA level, to associate it with expression quantity and phenotype, we conducted SNP/INDEL calling to each sample using Samtools program(http://samtools.sourceforge.net/) based on the assembled isogene sequences which were used as templates to BLAST the original sequencing reads, and filtered the original results with the criteria of QUAL>20 and the coverage>2.

Identification and functional enrichment analysis of differentially expressed genes(DEGs)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

We followed the method reporterd by Audic S. et al. (Audic and Claverie 1997) to detect the differential genes, DEseq software and edgeR(http://www.bioconductor.org/) were used to perform the differential gene expression analysis with the screening threshold of *q*-value<0.001 and |FoldChange|>2.

Gene functional enrichment analysis of DEGs was conducted with the method of GOseq (Young, Wakeeld et al. 2012). Gene ontology(GO) terms involving biological process(BP), cellular component(CC) and molecular function(MF). DEGs of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched from KEGG databse (Kanehisa 2008), pathways with $P \leqslant 0.05$ were considered significantly enriched by DEGs.

RESULTS

Illumina RNA-Seq and de novo assembly

In this study, 2 cDNA libraries of *Moschus berezovskii* (heart and musk gland) were constructed. These RNA samples were subjected to Illumina transcriptome sequencing. The raw reads count were 5.96 and 5.67 million, which contained 7.45 Gb and 7.08 Gb nucleotides (nt) from heart and musk gland, respectively. After removing low-quality sequences, a total of 5.93 and 5.64 million clean reads were obtained from heart and musk gland, respectively. The de novo transcriptomes were assembled from 239,383 transcripts with an average length of 769.11 bp (the shortest sequence was 201 bp and the longest one was 17,695 bp) (Fig. 1). Unigene was consisted of 208,730 contigs with the mean length of 599.31 bp (the shortest length was 201 bp and the longest length was 17,695 bp).

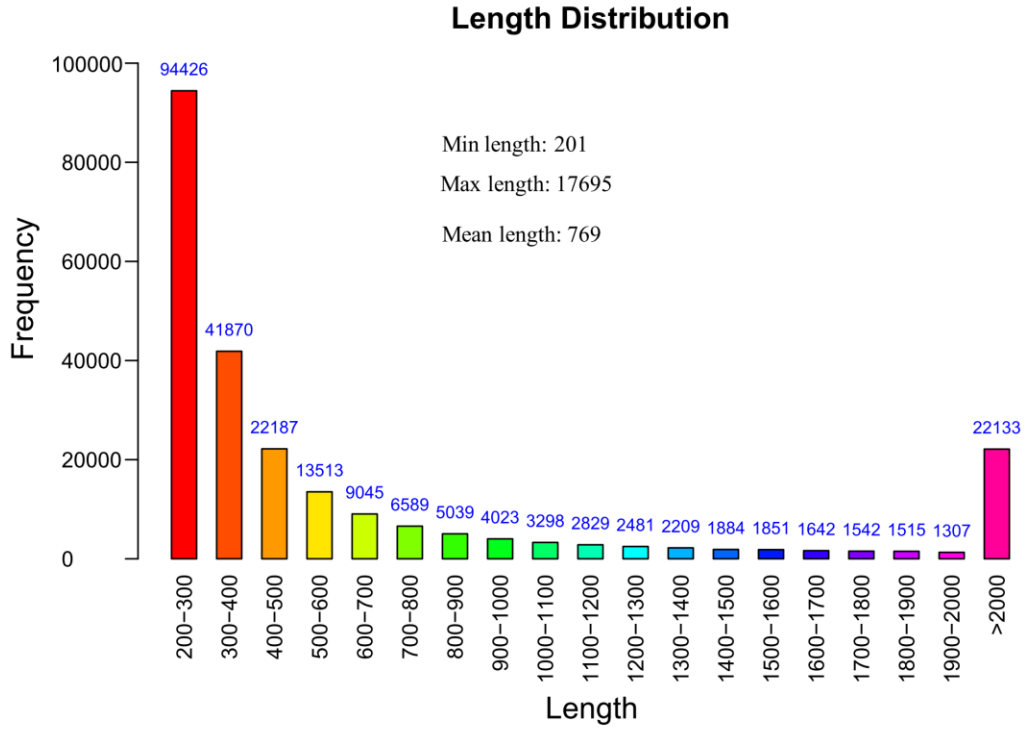

Figure 1 Length distribution of assembled transcripts of *Moschus berezovskii*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

Annotation of predicted proteins and homology search

Distinct gene sequences were searched using BLASTX against the NCBI nr database with a cut-off E-value of 10^{-5}. With this approach, 74,441(35.66%) matched to known NCBI nucleotide sequences (Nt) database, while 37,329 (17.88%) unigenes matched to known NCBI non-redundant protein sequences (Nr) database. Unigenes annotated in GO and KEGG were 31,039 (14.87%) and 11,782 (5.64%) respectively (Fig. S1).

The homology search of *Moschus berezovskii* unigenes were conducted using BLASTn against NCBI non-redundancy protein database(Nr) with a cut-off E-value of 10^{-5} (with the similarity more than 90% and the coverage more than 80%). A total of 37,329 genes of top 20 abundance species displayed matching in NCBI database, and some genes were homologous to more than one species. The highest similarity was seen with *Bos taurus* (29.38%), followed by *Gallus gallus* (11.77%), *Ovis aries*(11.69%) and *Bos grunniens mutus* (10.27%) (Table S1).

Clusters of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways annotation

According to GO (Gene Ontology) classification, 150,563 genes were annotated in 23 biological process categories, 114,576 genes were annotated in 18 cellular component categories and 41,288 genes were annotated in 20 molecular function categories. In biological process ontology, the transcripts were primarily distributed in cellular process (11.13%); cell and cell part accounted for most part (both were 11.31%) in the cellular component ontology; and binding was the most abundant in molecular function ontology. We noticed a high percentage of genes from the metabolic process and organelle categories. Few genes were found in the categories of morphogen activity, metallochaperone activity, protein tag, chemoattractant activity, translation regulator activity, chemorepellent activity and nutrient reservoir activity. Moreover, metabolic pathways involved in musk secretion were annotated in GO annotation in biological process such as negative regulation of aldosterone biosynthetic process (GO:0032348), positive regulation of aldosterone biosynthetic process (GO:0032349), monoterpenoid metabolic process(GO:0032348), flavone metabolic process (GO:0051552), flavonoid metabolic process (GO:0009812) and flavonol 3-sulfotransferase activity (GO:0047894) in molecular function category.

We also mapped the annotated unigenes to the reference pathways in KEGG database, in total, 5.64% were annotated in KEGG pathways at level 1, among them, there were 162 KOs (KEGG Orthology, which represent the catalogs of proteins or enzymes, each KO contains proteins functioning similar in the same pathway) involved in metabolic pathwys of terpenoid, steriod, steriod hormone, aldosterone-regulated sodium reaborption, ketone bodies, such as hydroxymethylglutaryl-CoA reductase, hydroxymethylglutaryl-CoA synthase, isopentenyl-diphosphate delta-isomerase, protein-S-isoprenylcysteine, prenyl protein peptidase, diphosphomevalonate decarboxylase, mevalonate kinase, prenylcysteine oxidase/farnesylcysteine lyase, protein farnesyltransferase subunit beta,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

phosphomevalonate kinase, sterol regulatory element-binding transcription factor 1, sterol O-acyltransferase, cholesterol mono-oxygenase, sterol 14-demethylase, 3-oxo-5-alpha-steroid 4-dehydrogenase 3. Additionally, 332 pathways were annotated in KEGG pathway at level 2, and the related metabolism of musk compounds pathways included terpenoid backbone biosynthesis (ko00900), steroid hormone biosynthesis (ko00140), synthesis and degradation of ketone bodies (ko00072), tropane (ko00960), triterpenoid biosynthesis (ko00909), ovarian steroidogenesis (ko04913), flavone and flavonol biosynthesis (ko00944).

SNP analysis

Based on approaches depicted in methods, we identified 44,470 SNPs in heart and 82,921 SNPs in musk gland, respectively (Fig. 2 A and B). C:G->T:A was the leading mutation type, followed by T:A->C:G, C:G->G:C, C:G->A:T, T:A->G:C and T:A->A:T. It was noted that, mutation numbers in musk gland were greater than heart in each mutation type (Fig. 2 C and D).

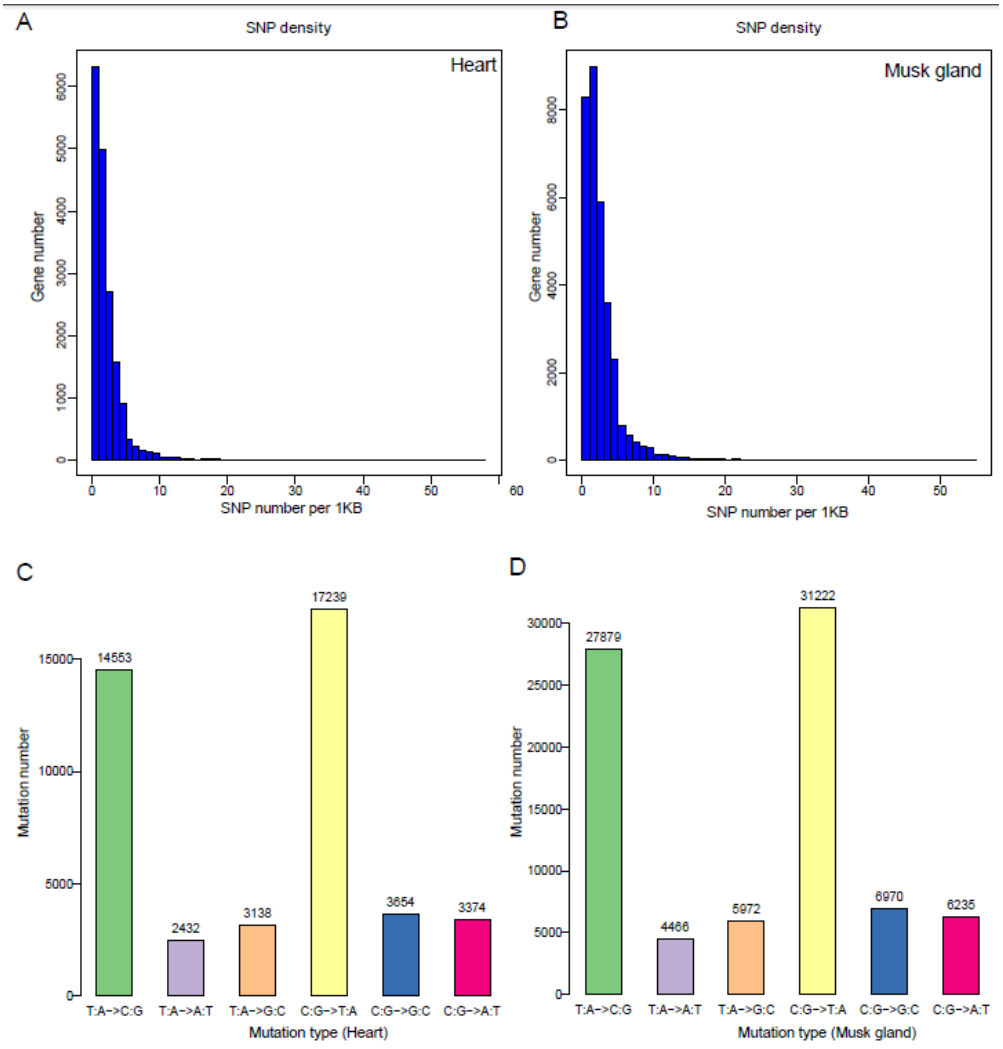

Figure 2 SNP density (panel A and B) and mutation spectrum (panel C and D) of heart and musk gland of *Moschus berezovskii*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

Identification of DEGs

To explore the differences among the transcriptomes between heart and musk gland, we performed identification of DEGs by comparing the transcripts data of heart and musk gland, a total of 8,986 genes were identified to be differentially expressed with the threshold of *P*-value <0.001 and the log2 (foldchange) >2. Out of these, 6,068 genes were up-regulated and 2,918 genes were down-regulated (Fig. 3 and Fig. 4). 25.44% (2,282 out of 8,969 DEGs) were specific expressed in musk gland, and 6.60% (592 out of 8,969 DEGs) were specific expressed in heart. Seventy-one DEGs which involved in musk formation and secretion were screened out, among them, 56 and 15 DEGs were up- and down-regulated, respectively. There were 6 genes (c33134_g1, c46915_g1, c99344_g1, c151678_g1, c125108_g1, c182615_g1) specific expressed in musk gland, which were functional enriched in pathways of steroid biosynthesis, aldosterone-regulated sodium reabsorption, steroid hormone biosynthesis, ovarian steroidogenesis and oocyte meiosis. It was noted that there was no specific expressed gene in heart among these 71 DEGs. There were 9 DEGs were enriched in more than 1 pathways, it might hint that these genes played significant roles in musk formation and secretion.

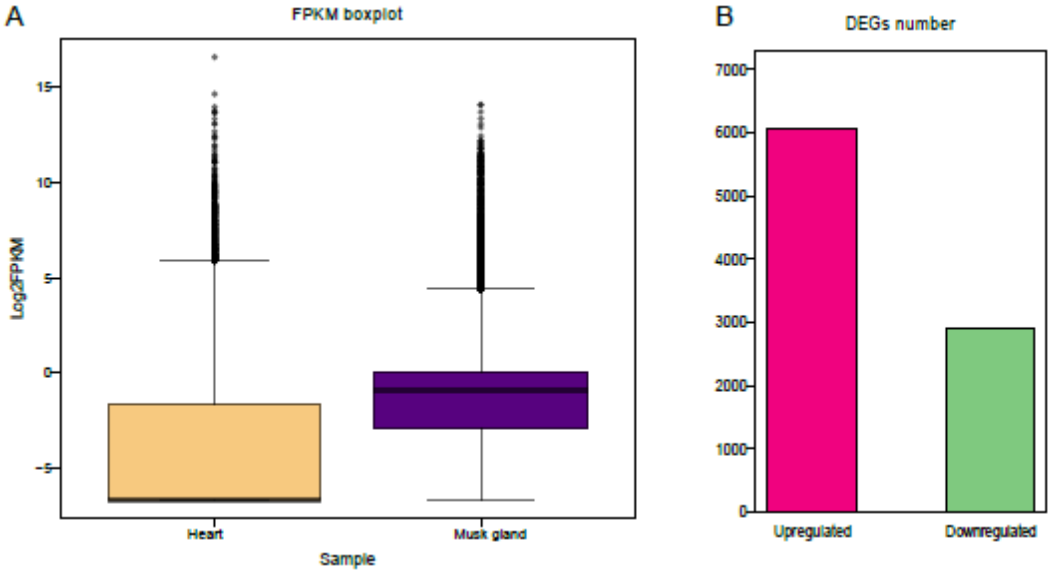

Figure 3 FPKM boxplot and DEGs of heart and musk gland of *Moschus berezovskii*. (A) Log2(FPKM) values of two different tissues. Tentacles express the range of maximum and minimum value of expression quantity, the boxes express 25%-75% of log2(FPKM), and the black lines in the box express the median. (B) The up- and down-regulated DEGs identified between heart and musk gland.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

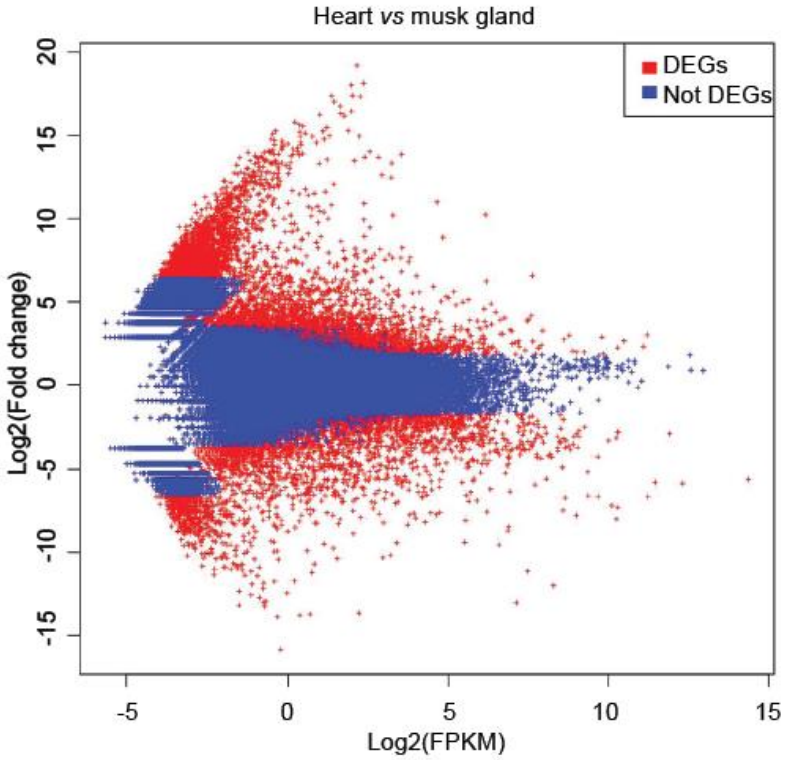

Figure 4 DEGs of heart and musk gland of *Moschus berezovskii*

Functional enrichment analysis of DEGs

To further elucidate the different functions of DEGs, we examined whether these DEGs were enriched for specific biological process(BP), molecular function(MF) and cellular component (CC) using the GOseq package. There were 1,888 (BP), 303 (MF) and 106 (CC) DEGs enriched on GO terms with the threshold of $P \leqslant 0.05$. The top 5 pathways with the largest number of DEGs were as follows: multicellular organismal process, single-organism process, single-multicellular organism process, single-organism cellular process, organ development in biological process; extracellular region, myofibril, contractile fiber, sarcomere, contractile fiber part in cellular component; structural constituent of muscle, transmembrane transporter activity, substrate-specific transmembrane transporter activity, transporter activity, calcium ion binding in molecular function. A total of 2,793 genes (1,968 and 825 up- and down-regulated DEGs, respectively) were differentially expressed in single-organism process and 2,453 genes (1,711 and 742 DEGs were up- and down-regulated, respectively) were differentially expressed in single-organism cellular process (Fig. 5). We screened 60 terms correlated with musk metobolism, most of them were up-regulated in musk gland compared to heart excepted 9 terms which were seen down-regulated such as negative regulation of aldosterone metobolic process, negative regulation of aldosterone biosynthetic process, ketone biosynthetic process, cellular ketone metabolic process.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

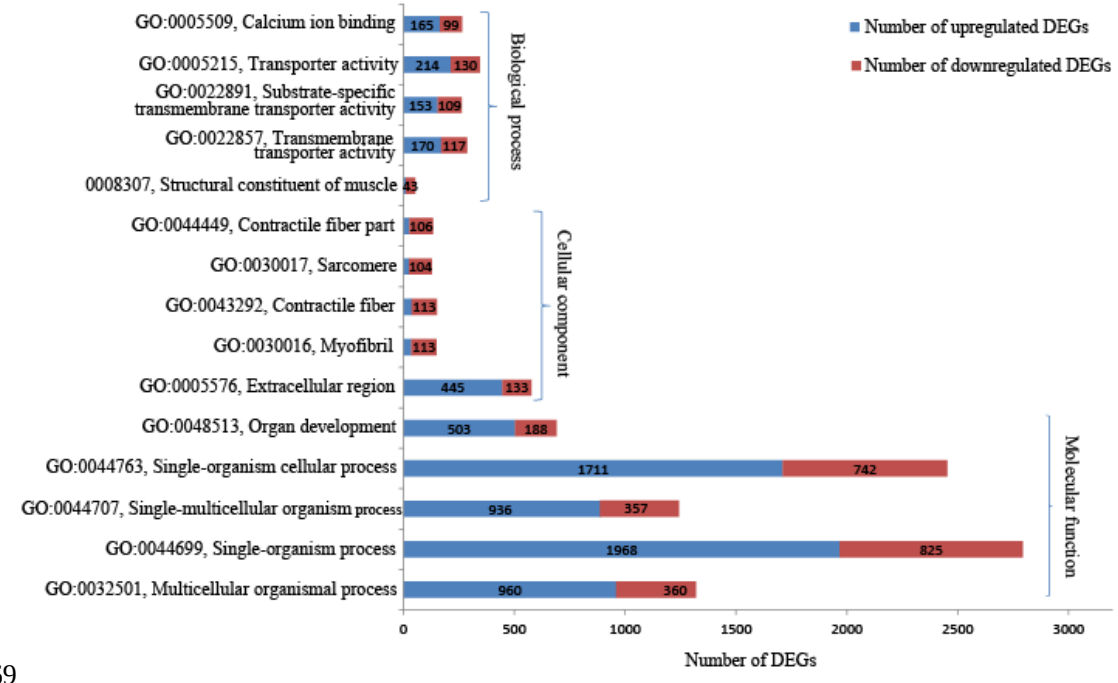

Figure 5 TOP 5 pathways with the largest number of DEGs enriched for up-(blue bars) and down-regulated(red bars) of *Moschus berezovskii*

There were 310 pathways enriched in KEGG functional annotation analysis of DEGs in heart and musk gland. We screened 10 pathways involved in musk compounds among all the pathways. Compared with heart and musk gland, 5 of them were up-regulated, 3 pathways were generally up-regulated including ovarian steroidogenesis (13/3), steroid hormone biosynthesis (14/1), aldosterone-regulated sodium reabsorption (10/3), and 2 pathways were down-regulated. However, only pathways of ovarian steroidogenesis (13/3) and steroid hormone biosynthesis(14/1) were significant different enriched in DEGs ($P \leqslant 0.05$) (Fig. 6).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

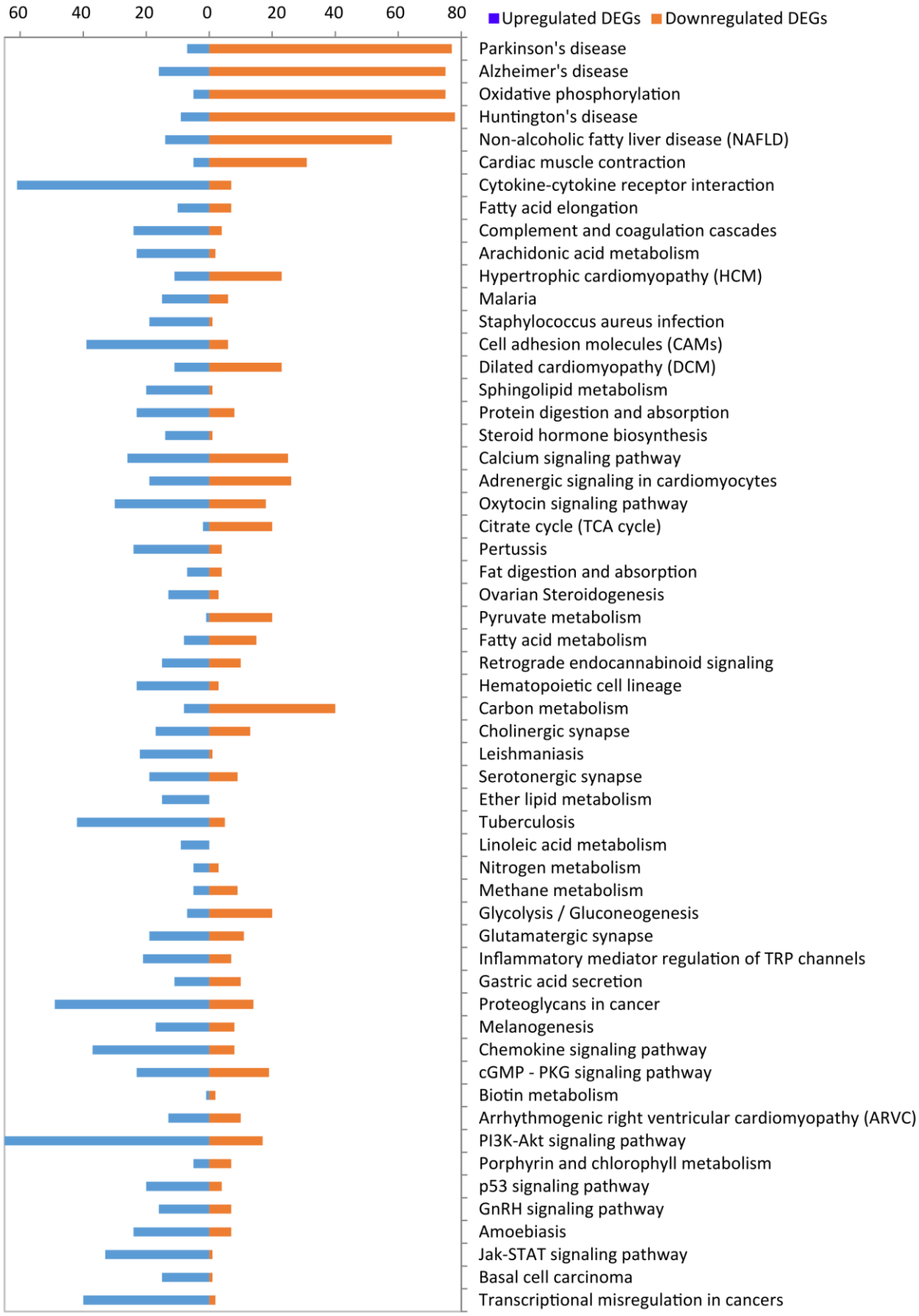

Figure 6 The KEGG pathways enriched for the differentiated expression genes between the heart and musk gland of *Moschus berezovskii*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

DEGs involved in musk secretion

To illuminate the mechanism of musk secretion in male Chinese forest musk deer, we have to form a comprehensive understanding of the chemical compound of musk, the metabolic processes of key chemical compounds, the corresponding pathways and genes involved in these processes. Therefore we screened 66 genes from all the annotated genes, 35 of them were novel genes. Here we list some important candidate genes which were closely involved in related pathways.

Firstly, genes involved in metabolism of steroid including *DHCR7*, *DHCR24*, *NSDHL* and *CYP3A5*, the first 3 genes were annotated in steroid biosynthesis, and *CYP3A5* was annotated in steroid hormone biosynthesis. *DHCR7* encodes an enzyme which catalyzes the conversion of 7-dehydrocholesterol to cholesterol. *DHCR24* encodes a flavin adenine dinucleotide (FAD)-dependent oxidoreductase which catalyzes the reduction of the delta-24 double bond of sterol intermediates during cholesterol biosynthesis. The protein contains a leader sequence that directs it to the endoplasmic reticulum membrane. The protein encoded by *NSDHL* is localized in the endoplasmic reticulum and is involved in cholesterol biosynthesis. *CYP3A5* encodes cytochrome P450 proteins which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. The encoded protein metabolizes drugs as well as the steroid hormones testosterone and progesterone.

Furthermore, genes participated terpenoid metabolism including *FDFT1* and *FDPS*. The former was annotated in sesquiterpenoid and triterpenoid biosynthesis, the encoded protein is the first specific enzyme in cholesterol biosynthesis, catalyzing the dimerization of two molecules of farnesyl diphosphate in a two-step reaction to form squalene. The latter was annotated in terpenoid backbone biosynthesis which encodes an enzyme that catalyzes the production of geranyl pyrophosphate and farnesyl pyrophosphate from isopentenyl pyrophosphate and dimethylallyl pyrophosphate. The resulting product, farnesyl pyrophosphate, is a key intermediate in cholesterol and sterol biosynthesis.

Besides, there was 1 gene (*HMGCL*) annotated in synthesis and degradation of ketone bodies, the protein encoded by this gene belongs to the HMG-CoA lyase family. It is a mitochondrial enzyme that catalyzes the final step of leucine degradation and plays a key role in ketone body formation. Mutations in this gene are associated with HMG-CoA lyase deficiency. Alternatively spliced transcript variants encoding different isoforms have been found for this gene

DISCUSSION

Next-generation sequencing (RNA-Seq) is an efficient method for investigating gene expression patterns, especially in non-model species that do not have sequenced genomes(Ockendon, O'Connell et al. 2015). With this approach, our primary goal was to understand characteristics of transcriptome in heart and musk gland of Chinese forest musk deer, identify DEGs between them. In this study, although we have used

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

highly advanced Illumina HiSeq 2500 platform with paired-end 125 bp to produce 99.58% and 99.51% high-quality sequences of heart and musk gland, respectively, yet only 39.3% (82,070/20,8730) unigenes have homologous matched to the entries in NCBI database with the cut-off E-value of 10^{-5}, and merely 14.87% can be annotated to one or more GO terms by GO analysis, which indicated that a large number of transcripts of Chinese forest musk deer were either non-coding or homologous with genes that did not have any GO term. As one unigene could align to more than one database, all the annotated unigenes in the whole databases (NR, NT, GO, KEGG, CDD, PFAM, TrEMBL, KOG and Swiss-Prot) comprised 18.02% of the total number of unigenes (Fig. S1 and Table S1), which should be considered adequate taking into account the lack of sequencing data on *Moschus*.

For the sake of evaluating the completeness of our transcriptome library and the effectiveness of our annotation process further, we searched the annotated sequences for genes associated with GO classification and KEGG pathways. In total, we assigned 31,039 unigenes to 61 GO terms, and the functional classification of these genes according to GO annotation identified genes associated to biological process, cellular component and molecular function categories. Activities related to musk compounds were annotated in GO annotation in molecular function and molecular classification, such as steroid dehydrogenase, steroid binding, steroid metabolic process, cholesterol metabolic process. Similarly, we assigned 208,730 unigenes to 332 KEGG pathways. Pathways involved in metabolic and synthetic activity of musk were annotated in KEGG annotation, i.e. flavone and flavonol biosynthesis, and terpenoid backbone biosynthesis which were coincident with the study by Li et al.(Li, Chen et al. 2016). It hinted that activities of steroid compounds (cholestanol, cholesterol, and a number of the androstane derivatives) were active in musk gland of forest musk deer. Furthermore, 46 genes were annotated in metabolism of terpenoids and polyketides pathways, which elucidated that it played an important role in formation of muscone.

To further explore the genes that showed differential expression patterns upon comparing the heart and musk gland, a total of 8,986 DEGs were detected, 6,068 of those were up-regulated and 2,918 were down-regulated. Genes possessing similar expression patterns potentially have a functional correlation, so we performed GO functional enrichment and KEGG pathway enrichment analysis to screen genes functionally correlated with musk secretion. Through GO functional enrichment analysis, the DEGs between heart and musk gland were categorized into 3 main categories including biological process, cellular component and molecular function (Fig. 5). We screened 60 GO terms which were correlated in metabolic pathways of musk compounds, the most terms were annotated in biological process such as aldosterone metabolic process, flavone metabolic process, aldosterone biosynthetic process, and terpenoid biosynthetic process (Table S2). Similarily, 10 pathways involved in metabolism of musk compounds were screened in KEGG enrichment analysis. It was noted that only pathways of ovarian steroidogenesis and steroid hormone biosynthesis were significant different enriched in DEGs(P<0.05). We identified several candidate genes such as *DHCR7*, *DHCR24*, *NSDHL*, *CYP3A5*,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

FDFT1, *FDPS* and *HMGCL* which were closely involved in metobolism of steroid, terpenoid and ketone bodies which were key compounds of musk.

Previous studies have shown that genes involved in reproduction regulation pathways such as steriod hormone biosynthesis, oocyte meiosis, steriod biosynthesis, GnRH signaling pathways were dominated and differentially expressed in Huoyan goose ovaries between the laying period and ceased period(Luan, Liu et al. 2014). This differential expression pattern is also the case in hens with different laying rates(Chen, Shiue et al. 2007). Examination of steriod production and steriod signaling was implemented for the sake of understanding the process of steriod-mediated oocyte maturation. Evaul(Evaul 2009) discovered that gonadotropin could induce steriod production, and then steriod-induced oocyte maturation came up. Broadly speaking, steriod metabolism plays a significant part in manipulating reproduction and has a close relationship with genital gland, in other words, metabolism of these compounds were involved in reproduction. Similarily, musk gland, which is located between naval and genital in male Chinese forest musk deer, is a representative sexual character of this species, and produces chemical compounds such as muscone and cholesterol which are primary composition of musk to manipulate the composition, formation and secretion of musk.

This study reported the transcriptomes of Chinese forest musk deer for the first time, method adopted was de novo assembly as a non-model species, united with unigenes annotation. We found many functional pathways related to steroid compounds metabolism, for example, steroid hormone biosynthesis, flavone and flavonol biosynthesis and steroid biosynthesis. The obtained transcriptome and DEGs data provide comprehensive gene expression information at the transcriptional level that could promote better understanding of the molecular genetic mechanisms underlying musk formation and secretion in Chinese forest musk deer.

CONCLUSION

In this study, the transcriptome of Chinese forest musk deer was sequenced using the Illumina Hiseq 2500 platform. A de novo assembly was conducted to identify candidate genes associated with musk secretion. It could come to the conclusion that pathways involved in musk secretion were up-regulated in musk gland compared with heart, such as steroid biosynthesis and terpenoid backbone biosynthesis whose metabolic productions were key components of musk, which was in accordance with the report of Li et al.(2016). Related candidate genes included *DHCR7*, *DHCR24*, *NSDHL*, *CYP3A5*, *FDFT1*, *FDPS* and *HMGCL* which were closely involved in metobolism of steroid, terpenoid and ketone bodies. These data offered clues to the mechanisms of musk secretion, further studies will incorporate RT-PCR and molecular experiments to investigate the specific role of transcription factors and verify candidate genes in Chinese forest musk deer in order to develop molecular tools for exploring the mechanism of musk secretion, seeking efficient and scientific means to protect and utilize *Moschus* resources, and finding some ways to increase the production and quantity of musk in genetic ways.

ACKNOWLEDGEMENTS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

We would like to thank Sangon Biotech for providing sequencing service and primary data analysis.

REFERENCES

Ana C, Stefan GT, Miguel GGJ, Javier T, Manuel T, Montserrat R. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 21(18): 3674-3676 DOI: 10.1093/bioinformatics/bti610.

Audic S, Claverie JM. 1997. The significance of digital gene expression profiles. *Genome Research.* 7(10): 986-995 DOI:10.1101/gr.7.10.986.

Bi SZ, Shen Y. 1986. Study on morphology and chemical communication mechanism in musk gland. *Chinese Journal of Zoology.* 02(04): 11-14.

Bi SZ, Shen Y, Zhu DX. 1984. Study on ultra-microstructure and musk secretion of musk gland in prosperous secreting stage. *Acta Theriologica Sinica*. 02(2): 81-85.

Cao XH, Zhou YD. 2007. Progress on anti-inflammatory effects of musk. *China pharmacy.* 18(21): 1662-1665.

Chen CF, Shiue YL, Yen CJ, Tang PC, Chang HC, Lee YP. 2007. Laying traits and underlying transcripts, expressed in the hypothalamus and pituitary gland, that were associated with egg production variability in chickens. *Theriogenology.* 68(9): 1305-1315 DOI:10.1016/j.theriogenology.2007.08.032.

Chen X. 2007. Studies on the genetic diversity of forest musk deer(*Moschus berezovskii*) and linkage analysis between the performance of musk productivity and AFLP markers. M. Scie. Thesis, Zhejiang University.

Deng FM. 1986. Domestication and grazing control of forest musk deer (*Moschus berezovskii*). *Chinese Journal of Wildlife*. 4: 35-37.

Evaul KE. 2009. Gonadotropin-induced steroidogenesis and downstream signals leading to oocyte maturation. D. Phil. Thesis, University of Texas Southwestern Medical Center at Dallas.

Feng QQ, Liu TJ. 2015. Progress on pharmacological activity of muscone. *Food and drug.* (3): 212-214.

Grabherr MG, Haas BJ, Moran Y, Levin JZ, Thompson DA, Ido A, Xian A, Lin F, Raktima R, Qian DZ. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.* 29(7): 644-652 DOI: 10.1038/nbt.1883.

Guan TL, Zeng B, Peng QK, Yue BS, Zou FD. 2009. Microsatellite analysis of the genetic structure of captive forest musk deer populations and its implication for conservation. *Biochemical Systematics and Ecology.* 37(3): 166-173 DOI: 10.1016/j.bse.2009.04.001.

Jia Y, Lin F, Hongkun Z, Yong Z, Jie C, Zeng JZ, Jing W, Shengting L, Ruiqiang L, Lars B. 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research.* 34(Web Server issue): W293–W297 DOI: 10.1093/nar/gkl031.

Kanehisa M. 2008. The KEGG Database. 'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247. *John Wiley & Sons, Ltd*. 247:91-103.

Li DY, Chen BL, Zhang L, Gaur U, Ma TY, Jie H, Zhao GJ, Wu N, Xu ZX, Xu HL, Yao YF, Lian T, Fan XL, Yang DY, Yang MY, Zhu Q, Satkoski TJ. 2016. The musk chemical composition and microbiota of Chinese forest musk deer males. *Scientific Report.* 6: 18975. DOI:10.1038/srep18975.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

Liu M, Yao B, Zhang H, Guo H, Hu D, Wang Q, Zhao Y. 2014. Identification of novel reference genes using sika deer antler transcriptome expression data and their validation for quantitative gene expression analysis. *Genes & Genomics.* 36(5): 573-582 DOI: 10.1007/s13258-014-0193-x.

Luan X, Liu D, Cao Z, Luo L, Liu M, Gao M, Zhang X. 2014. Transcriptome profiling identifies differentially expressed genes in Huoyan goose ovaries between the laying period and ceased period. *Plos One.* 9(9): e113211-e113211 DOI: 10.1371/journal.pone.0113211.

Maudhoo MD, Ren D, Gradnigo JS, Gibbs RM, Lubker AC, Moriyama EN, French JA, Jr RBN. 2014. De novo assembly of the common marmoset transcriptome from NextGen mRNA sequences. *Gigascience.* 3(1): 1-4 DOI: 10.1186/2047-217X-3-14.

Mortazavi A, Williams BA, Mccue K,chaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods.* 5(7): 621-628 DOI: 10.1038/nmeth.1226.

Ockendon LA. O'Connell S, Bush J, Monzón-Sandoval J, Barnes JH, Székely T, Hofmann HA, Dorus S, Urrutia AO. 2015. Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Molecular Ecology Resources* DOI: 10.1111/1755-0998. 12465.

Peng H, Liu S, Zou F, Zeng B, Yue B. 2009. Genetic diversity of captive forest musk deer (*Moschus berezovskii*) inferred from the mitochondrial DNA control region. *Animal Genetics.* 40(1): 65-72 DOI: 10.1111/j.1365-2052.2008.01805.x.

Seth SD, Mukhopadhyay AB, Prbhakar MC. 1973. Antihista-minic and spasmolytic effects of musk. *Japanese Journal of Pharmacology.* 23(5): 673-679.

Sheng HL. 1996. Protection and utilization of musk deer resources in China. *Chinese Journal of Wildlife.* 91: 10-12 DOI: 10.3732/ajb.1100292.

Strickler SR, Aureliano B, Mueller LA. 2012. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany.* 99(2): 257-266 DOI:10.3732/ajb.1100292.

Su B, WangYX, Wang QS. 2001. Mitochondrial DNA sequences imply Anhui musk deer a valid species in genus *Moschus*. *Zoological Research.* 22(3): 169-173.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology.* 17(7): 1636-1647 DOI: 10.1111/j.1365-294X.2008.03666.x.

Zhang YN, Jin JY, Jin R, Xia YH, Zhou JJ, DengJY, Dong SL. 2013. Differential expression patterns in chemosensory and non-chemosensory tissues of putative chemosensory genes identified by transcriptome analysis of insect pest the purple stem borer Sesamia inferens (Walker). *Plos One.* 8(7) DOI: 10.1371/journal.pone.0069715.

Young MD, Wakeeld MJ, Smyth GK, Oshlack A. 2012. goseq: Gene Ontology testing for RNA-seq datasets. 1-20.

Zhao Y, Yao. BJ, Zhang M, Wang SM, Zhang H and Xiao W. 2013. Comparative analysis of differentially expressed genes in Sika deer antler at different stages. *Molecular Biology Reports.* 40(2): 1665-1676 DOI 10.1007/sl 1003-012-2216-5.

Zhang Z, Deng Z, Li ZC. 1985. Domestication and trans-cultivation of forest musk deer(*Moschus berezovskii*). *Jorunal of Chinese Medicinal Materials.* (2): 14-15.

Zhao SS. 2009. Assement of genetic diversity in the captive forest musk deer(*Moschus berezovskii*) and linkage analysis between the performance of musk productivity and DNA molecular markers. D. Scie. Thesis, Zhejiang University.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

498

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2252v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016