Journal of Genetics, Vol. 96, No. 6, December 2017, pp. 1033–1040
https://doi.org/10.1007/s12041-017-0872-x

RESEARCH ARTICLE

Illumina-based *de novo* transcriptome sequencing and analysis of Chinese forest musk deer

ZHONGXIAN XU[1], HANG JIE[2], BINLONG CHEN[1], UMA GAUR[1], NAN WU[1], JIAN GAO[1], PINMING LI[2], GUIJUN ZHAO[2], DEJUN ZENG[2], MINGYAO YANG[1] and DIYAN LI[1]*

[1]*Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu 611130, People's Republic of China*
[2]*Laboratory of Medicinal Animal, Chongqing Institute of Medicinal Plant Cultivation, Nanchuan, Chongqing 408435, People's Republic of China*
*For correspondence. E-mail: diyanli@sicau.edu.cn.

Received 22 November 2016; revised 27 March 2017; accepted 28 April 2017; published online 18 December 2017

Abstract. The Chinese forest musk deer (*Moschus berezovskii* Flerov) is an endangered artiodactyl mammal. The musk secreted by sexually mature males is highly valued for alleged pharmaceutical properties and perfume manufacturing. However, the genomic and transcriptomic resources of musk deer remain deficiently represented and poorly understood. Next-generation sequencing technique is an efficient method for generating an enormous amount of sequence data that can represent a large number of genes and their expression levels. In the present study, we used Illumina HiSeq technology to perform *de novo* assembly of heart and musk gland transcriptomes from the Chinese forest musk deer. A total of 239,383 transcripts and 176,450 unigenes were obtained, of which 37,329 unigenes were matched to known sequences in the NCBI nonredundant protein (Nr) database; 31,039 unigenes were assigned to 61 GO terms, and 11,782 to 332 KEGG pathways. Additionally, 592 and 2282 differentially expressed genes were found to be specifically expressed in the heart and musk gland, respectively. The abundant transcriptomic data generated in the present report will provide a comprehensive sequence resource for Chinese forest musk deer as well as lay down a foundation which will help in accelerating genetic and functional genomics research in this species.

Keywords. Chinese forest musk deer; Illumina sequencing; *de novo* assembly; transcriptome dataset; *Moschus berezovskii*.

Introduction

Chinese forest musk deer is one of the seven species of *Moschus*, with entirely Asian present day distributions (the earliest known musk deer is an extinct European species from Oligocene deposits), and is famous for secretion of musk. It has been listed as an endangered species (Wang and Harris 2015), as well as a first class, 'key' wildlife species (Guan *et al*. 2009). Factors leading to these designations include a serious decline in population size caused by overexploitation, distribution shrinkage, habitat destruction and degradation. Musk, which is secreted by the musk gland of sexually mature males, has long been used in Chinese traditional medicine and perfume making (Sheng 1996; Su *et al*. 2001), because of its unique fragrance and its significant anti-inflammatory and antitumour roles, as well as its effects on the human central nervous and cardio–cerebral–vascular systems (Cao and Zhou 2007; Feng and Liu 2015). As a charismatic species, Chinese forest musk deer have been a 'hot spot' of ecology, domestication (Zhang *et al*. 1985; Deng 1986), and musk pharmacological (Seth *et al*. 1973; Feng and Liu 2015). In recent years, several research groups have studied the species' anatomy (Bi *et al*. 1984; Bi and Shen 1986), microsatellite sequences (Guan *et al*. 2009; Peng *et al*. 2009), mtDNA markers (Chen 2007; Zhao 2009), and microbiota (Li *et al*. 2016). Unfortunately, genomic

Zhongxian Xu and Hang Jie contributed equally to this work.

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12041-017-0872-x) contains supplementary material, which is available to authorized users.

and transcriptomic resources of the species have largely remained unexplored.

Next-generation sequencing technology is an important method for analysing transcriptomes, and has been extensively used to study a number of species ranging from yeast to human (Martin and Wang 2011). In the absence of an appropriate reference genome/transcriptome, as is often the case with nonmodel organisms, *de novo* assembly is the optimal choice for sequence assembly (Strickler *et al*. 2012; Ockendon *et al*. 2015). Previous studies have reported the transcriptome expression of Chinese sika deer antler in five developmental stages (Zhao *et al*. 2013; Liu *et al*. 2014), as a representative example in artiodactyl mammals.

We used the Illumina HiSeq platform to investigate the transcriptome of heart and musk gland tissues of Chinese forest musk deer. The resulting *de novo* assembly generated enormous (more than 56.00 million clean reads) gene fragments. Unigenes (the longest transcript of each gene) were obtained through similarity searches using the gene fragments against nine different databases. We also compared gene expression patterns between the heart and musk gland. This sequencing and annotation data will provide a valuable resource for further studies on Chinese forest musk deer.

Materials and methods

Sample collection, RNA isolation and sequencing

Captive Chinese forest musk deer were raised at the Chongqing Institute of Medicinal Plant Cultivation (Nanchuan, Chongqing, China). Heart and musk gland tissues were collected from single sexually mature unmated male individual after it had died of natural disaster (earthquake of 20 April 2013). Tissues were cut into small pieces and immediately placed in liquid nitrogen, and then stored at −80°C until RNA extraction. Total RNA was isolated from tissue samples using TRIzol reagent (Qiagen, Hilden, Netherlands) following the manufacturer's instructions. The RNA quality and quantity were measured with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, USA) and agarose gel electrophoresis. RNA samples having a RIN (RNA integrity number) >8 were used for constructing paired-end (PE) cDNA libraries. The PE cDNA libraries were sequenced using the Illumina HiSeq 2000 platform at Sangon Biotech (Shanghai, China) following manufacturer's protocols (Illumina, San Diego, USA), and PE reads of 125 bp or more for each library were obtained. Sequence files were generated in FASTQ format (sequence reads plus quality information in Phred format), and the raw sequence data was deposited in the National Center for Biotechnology Information (NCBI) under Bioproject ID: PRJNA291827 (heart) and PRJNA289641 (musk gland). The Illumina sequences were deposited in the SRA with accessions SRP061975 (heart) and SRP060734 (musk gland).

De novo assembly and annotation

Quality control and preprocessing of the raw sequence data were performed with FastQC software (ver. 0.11.5) (Andrews 2010, FastQC: A quality control tool for high throughput sequence data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Cutadapt software (ver. 1.2.1) (Martin 2011) was used for trimming adaptors, removing read-ends of less than 20 bp, reads containing unknown bases, and reads with minimum lengths less than 35 bp. Next, the PrinSeq program (ver. 0.20.3) (Schmieder and Edwards 2011) was used for removing duplicated reads with average qualities >Q20. All clean reads were pooled and subjected to transcriptome *de novo* assembly using the Trinity program (ver. 2.1.1) (Grabherr *et al*. 2011; Haas *et al*. 2013), with a minimum contig length of 200 (k = 200). To evaluate for contamination in the assembled transcriptome, 100,000 clean reads were randomly selected and aligned to sequences in the NCBI nucleotide (Nt) database using BLASTn (ver. 2.3.0) with E-value ≤ 10^{-10}. We then searched all the translations from our unigene sequences using BLASTx (E-value < 0.00001) against seven other databases: NCBI's non-redundant protein sequences (Nr), UniProt's manually annotated Swiss-Prot database (Swiss-Prot) along with those machine annotated translations from EMBL not in Swiss-Prot (TrEMBL), the European Bioinformatics Institute's protein family hidden Markov model database (Pfam), NCBI's conserved domain database of position-specific score matrices (CDD), the Gene Ontology Consortium's gene attributes relational database (GO), the Kyoto encyclopedia of genes and genomes (KEGG), and the Joint Genome Institute's euKaryotic Orthologous Groups (KOG) tool. Subsequently, we performed GO annotation of the unigenes using Blast2GO (Conesa *et al*. 2005) through the BLASTx program, and KEGG Orthology (KO) terms and KEGG pathway annotation were conducted using KAAS (KEGG Automatic Annotation Server) (Moriya *et al*. 2007) to identify the genes associated with various pathways.

Analysis of differentially expressed genes

Gene expression levels were expressed as fragments per kilobase of transcript per million mapped reads (FPKM) (Mortazavi *et al*. 2008) using Rsem software (ver. 1.2.31) (Li and Dewey 2011). Differential gene expression analysis was undertaken with unreplicated samples following established methods (Audic and Claverie 1997). DESeq (Anders and Huber 2013) was used to identify differentially expressed genes (DEGs), and the significance of gene expression differences were assessed using a screening threshold of false discovery rate (FDR) ≤0.001 and

fold change >2. DEG functional enrichment analysis of GO term (Harris *et al*. 2004) was performed with GOseq (Young *et al*. 2012). KEGG pathway (Kanehisa 2008) assignment for DEGs and pathway enrichment analysis were performed using KOBAS (ver. 2.0) (Xie *et al*. 2011). GO terms and KEGG pathways with adjusted *P* values <0.05 were identified as significantly enriched.

Results

De novo assembly and annotation of Chinese forest musk deer transcriptome

We constructed two cDNA libraries from Chinese forest musk deer (heart and musk gland). The raw read count was 59.60 and 56.67 million, containing a total of 7.45 and 7.08 Gb in heart and musk gland, respectively. After removing adaptors, ambiguous sequences, and low-quality reads, a total of 59.34 and 56.39 million clean reads were generated. After transcription, the *de novo* assembly generated 239,383 transcripts with an average length of 769.11 bp (201 bp to 17,695 bp range) (figure 1a). Of these, 176,450 were assigned unigene status, with a mean length of 599.31 bp (201 bp to 17,695 bp range).

We conducted similarity searches using our unigenes against nine databases to annotate the assembled unigenes. As a result, 74,441 (42.19%), 37,329 (21.16%), 31,039 (17.59%), and 11,782 (6.68%) unigenes were matched to Nt, Nr, GO and KEGG database entries, respectively (figure 1b). Unigenes with similarities more than 30%, and E-values <0.00001 were annotated in detail. A species distribution analysis of our BLASTx search results against the Nr database showed that our unigenes were most represented in *Bos taurus* (29.38%), followed by *Gallus gallus* (11.77%), *Ovis aries* (11.69%) and *Bos grunniens mutus* (10.27%) (figure 2a), i.e. many of the genes we analysed in the Chinese forest musk deer were also found in other artiodactyl mammals. Unigenes aligned to Nt database sequences were used to evaluate the contamination of the RNA samples. Of the top nine hits, *Moschus chrysogaster* accounted for most; the seven other top hits were other *Moschus* species, indicating lack of RNA contamination (figure 2b).

Functional annotation of unigenes

According to GO clustering, 150,563 (85.33%) of our unigenes were annotated by 23 biological process (BP) categories, 114,576 (64.93%) were involved in 18 cellular component (CC) categories, and 41,288 (23.40%) participated in 20 molecular function (MF) categories. The most abundant terms were distributed in 'cellular process' (23,227; 15.43%) in the BP categories, 'cell' and 'cell part' (23,616; 20.61%) in the CC categories, and 'binding' (18,750; 45.41%) in the MF categories (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Additionally, we noticed that a high percentage of genes were related to 'metabolic process' and 'organelle' categories, and only a high percentage of genes were involved in the categories 'morphogen activity', 'metallochaperone activity', 'protein tag', and 'nutrient reservoir activity'. We categorized 16,108 (9.13%) unigenes into 32 subgroups containing five KEGG classification types (table 1 in electronic supplementary material), in order of prevalence: 'organismal system' (5065; 31.44%), 'metabolism' (3739; 23.21%), 'environmental information processing' (3432; 21.31%), 'cellular process' (1980; 12.29%), and 'genetic information processing' (1892; 11.75%). The predominant subgroups were 'signal transduction' (2762; 17.15%), 'immune system' (1382; 8.58%), and 'endocrine system' (1172; 7.28%).

Identification of DEGs and functional enrichment analysis

Of the total 176,450 expressed unigenes, 72,724 (41.22%) coexpressed in both tissues, while 17,337 (9.83%) and 86,389 (48.95%) unigenes were specifically expressed in heart and musk gland, respectively (table 2 in electronic supplementary material). The average median expression levels (Log_2 FPKM) of heart and musk gland were -6.95 and -1.02, respectively (figure 3a). A total of 8986 genes were identified as differentially expressed of which nearly 25.40% (2282 of 8986 DEGs) were specifically expressed in the musk gland, and 6.59% (592 of 8986 DEGs) were specifically expressed in the heart. Six thousand sixty eight genes were upregulated and 2918 genes were downregulated in the musk gland (figure 3b), as compared with the same genes' regulation in the heart.

To further determine different DEG functions, we examined all 8986 DEGs for specific BP, MF, and CC term enrichment using the GOseq package (Young *et al*. 2012). A total of 2297 terms were obtained in the DEGs; of them 1888 (BP), 303 (MF), and 106 (CC) were enriched with GO terms with *P* values <0.05 (table 3 in electronic supplementary material). We further screened 17 terms associated musk metabolism most of which were upregulated in the musk gland as compared with the same genes' regulation in the heart, except for three (ubiquinone metabolic process, ubiquinone biosynthetic process and mitochondrial electron transport, NADH to ubiquinone), which were specifically expressed in the heart (table 3 in electronic supplementary material). Nine out of 17 DEGs were involved in 'steroid metabolism' followed by 'ubiquinone' (4), 'ketone' (2) and 'terpenoid' (2).

Fifty-six significantly enriched pathways ($P < 0.05$) were identified in the KEGG pathway classification (table 4 in electronic supplementary material). The top most three enriched pathways were Alzheimer's disease (16 upregulated and 75 downregulated genes, as compared with the

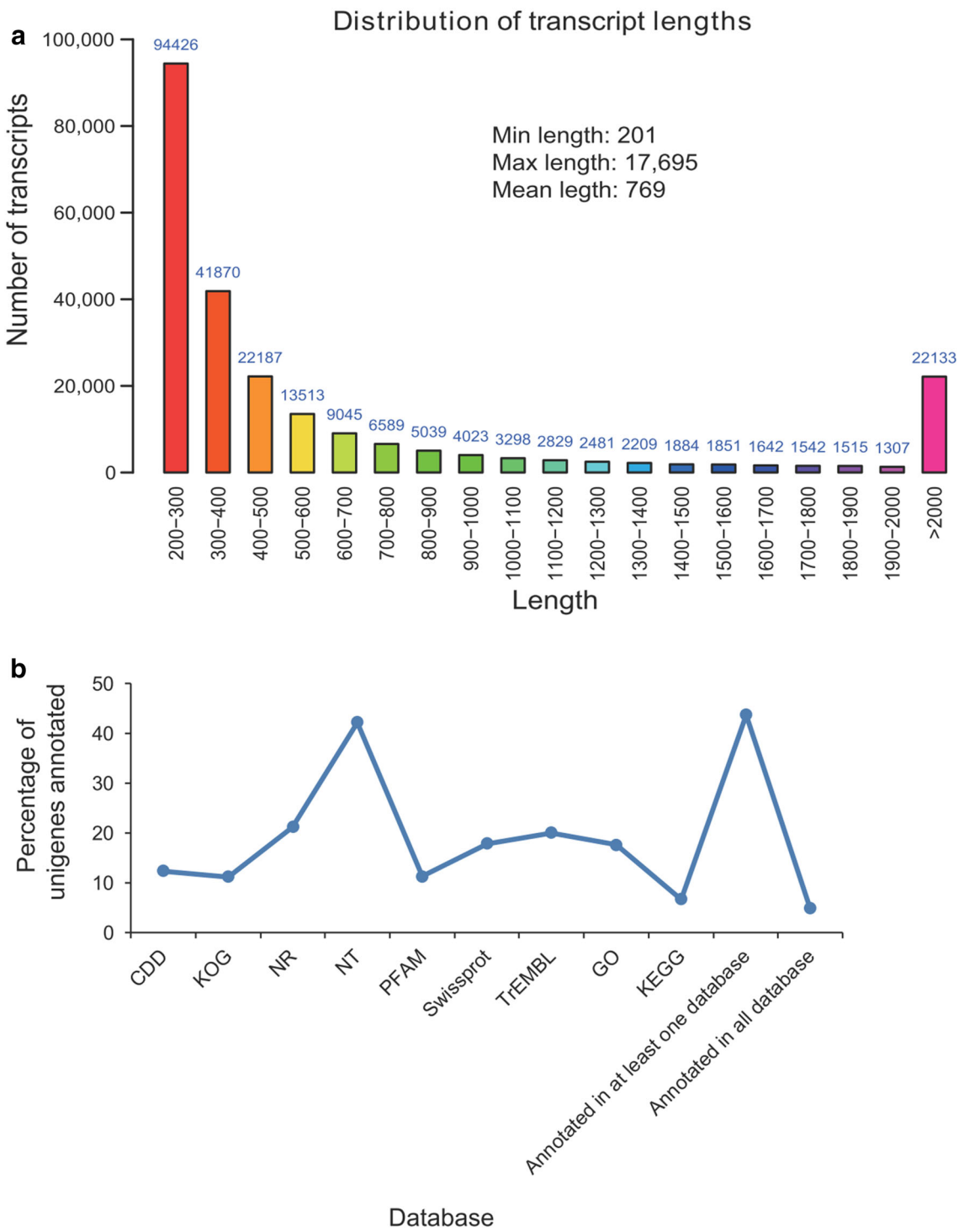

Figure 1. (a) Length distribution of assembled transcripts and (b) percentage of unigenes annotated from nine databases.

same genes' regulation in the heart), Huntington's disease (nine upregulated and 78 downregulated genes, compared to the heart), Parkinson's disease (seven upregulated and 77 downregulated genes, compared to the heart). There were 14 upregulated DEGs involved with steroid hormone biosynthesis in the musk gland, as compared to the counterpart genes in the heart (one downregulated DEG, compared to the heart), and 16 of 23 DEGs involved with the GnRH signalling pathway were upregulated in the musk gland (seven downregulated DEGs), compared to the heart. This information suggests that pathways involved in musk production are expressed at a very low

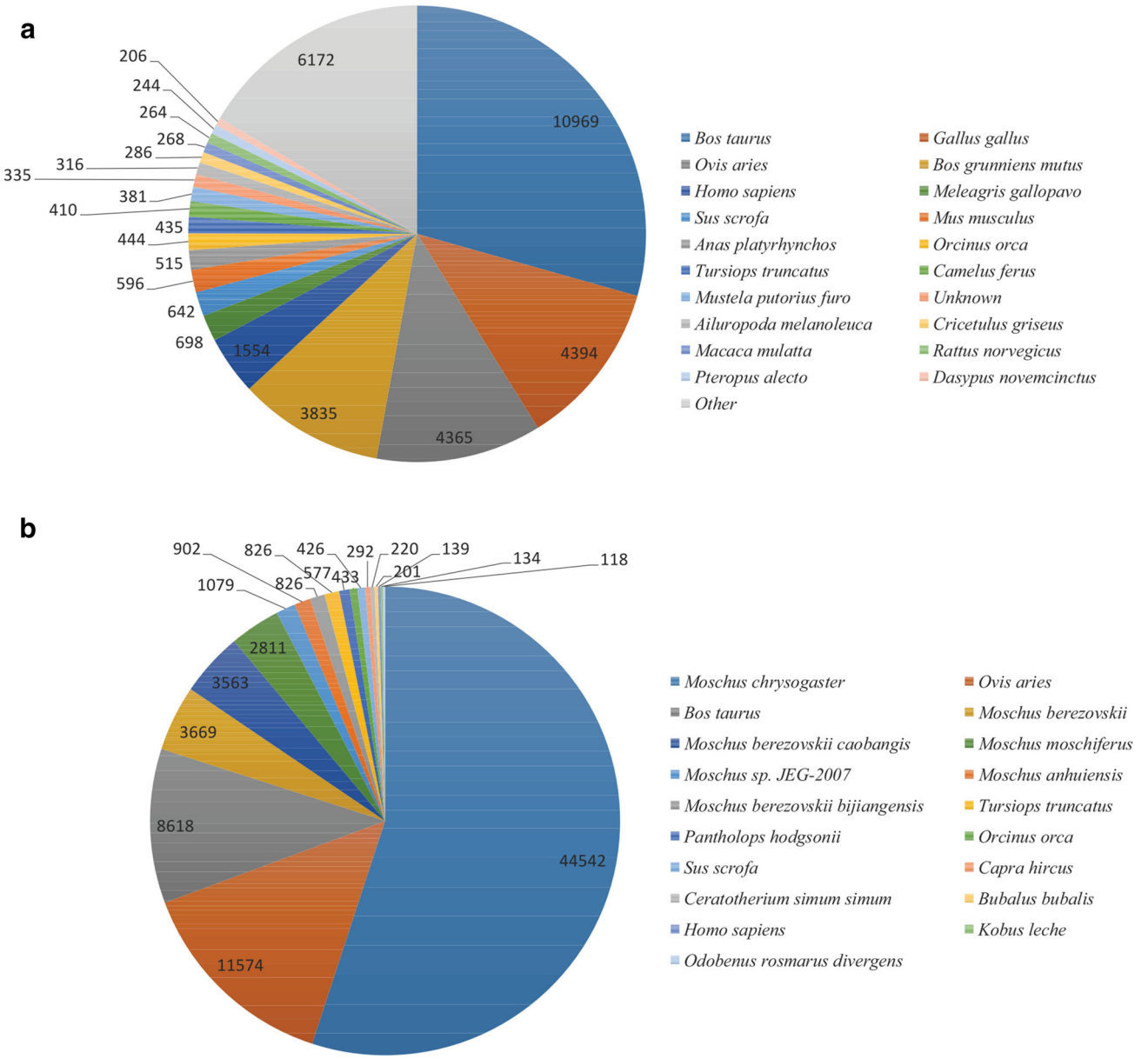

Figure 2. Similarity search characteristics of assembled unigenes. (a) Species distribution of the best BLASTx hit for each unigene against the Nr database. (b) Species distribution of the best BLASTn hit for each unigene against the Nt database.

level, but could play vital roles in musk metabolism. We discovered that genes such as *STS*, *UGT1A3*, *HSD17B7*, *CYP1B1* and *SRD5A1* were annotated by these two pathways; therefore, these genes might encode proteins participating in the regulation of musk production.

Discussion

Illumina HiSeq (RNA-Seq) is an efficient method for investigating gene expression patterns, especially in non-model species that do not have sequenced genomes (Ockendon *et al*. 2015). With this approach, our primary goals were to report transcriptome datasets from the heart and musk gland of Chinese forest musk deer, identify DEGs between them, and provide genetic resources for further molecular biological study in this species.

In total, 176,450 unigenes were found in our analysis. Unigene annotations were derived by mapping against nine databases (Nt, Nr, Swiss-Prot, TrEMBL, PFAM, CDD, GO, KEGG and KOG). This comprised 4.87% (8586) of the total number of unigenes we generated. The small percentage could be explained by the lack of available Cervidae genomic data. We found that most of our annotated unigenes were associated with *B. taurus* data, secondly to *G. gallus,* followed by *O. aries*, and

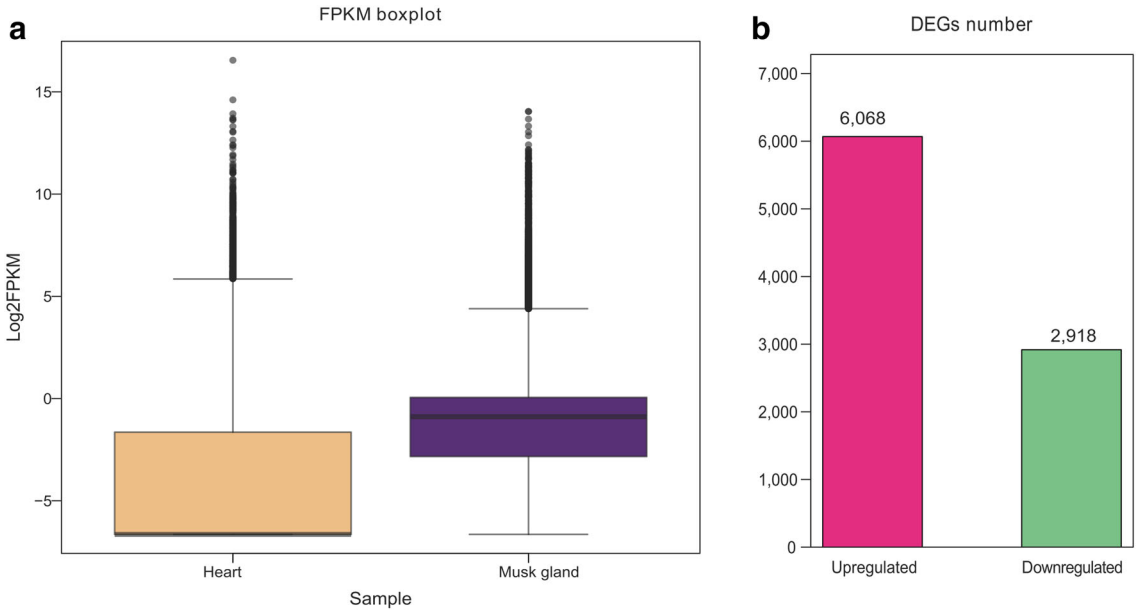

Figure 3. Boxplot of FPKM values and the number of differentially expressed genes (DEGs) by comparison between the heart and musk gland of the Chinese forest musk deer. (a) Boxplot of Log_2 FPKM values of two different tissues; the black lines represent median values. (b) Number of upregulated and downregulated DEGs identified in the musk gland, compared with the same genes' regulation in the heart.

then *B. g. mutus*, using the Nr database. This suggests that the Chinese forest musk deer is more closely related to *B. taurus* than *O. aries*, but is by no means conclusive. Most unigenes in our study were annotated by bird data rather than sheep (artiodactyla), probably due to the abundance of chicken genomic data (3,516,449 of our annotated entries) versus sheep (318,957 of our annotated entries). The Chinese forest musk deer expression patterns that we ascertained from our transcriptome dataset were similar to other artiodactyl mammal representatives (Zhao *et al*. 2013; Chen *et al*. 2015; Deng *et al*. 2016; Jia *et al*. 2016). In particular, the patterns were consistent with sika deer, of which transcriptome datasets from different development stages and 10 tissue types have been reported, including heart and testes (Jia *et al*. 2016). These results showed the number of testes-enriched unigenes to be ~15-fold enriched, compared with the heart. The number of musk gland specific unigenes was nearly five-fold enriched, compared with the heart, in our analysis. The musk gland, which is located between the naval and the genitals in male Chinese forest musk deer is a representative sexual characteristic in this species. We, thus, made the possible assumption that the musk gland might have similar expression patterns as testes, to a certain extent. Further comparison between the musk gland and testes of Chinese musk deer is needed to justify this assumption.

Among all the unigenes we annotated, we assigned 31,039 to 61 different GO terms. Analogous results (expression patterns of GO terms) have also been reported in sheep (Chen *et al*. 2015) and sika deer (Jia *et al*. 2016). Activities related to musk production, annotated by GO terms in the MF and BP categories include the following: steroid dehydrogenase, steroid binding, steroid metabolic process, and cholesterol metabolic process. Similarly, we assigned 11,782 unigenes to 332 KEGG pathways. Our results were consistent with the swamp buffalo, with its pooled RNA sample dataset from 11 different tissues (Deng *et al*. 2016). Deng assigned 331 KEGG pathways to his samples, and the patterns of main assigned categories that he identified were much the same as in our study, with the largest category including human diseases. However, the second most prevalent main category in our Chinese forest musk deer analysis was metabolism, rather than organismal systems. This might indicate that the musk gland of Chinese forest musk deer expresses more pathways (compared to swamp buffalo) involved in metabolism than other categories, besides human diseases. Pathways associated with the metabolism of musk annotated by the KEGG pathway assignment, e.g. flavone and flavonol biosynthesis and terpenoid backbone biosynthesis in our analysis were coincident with a study by Li *et al*. (2016) which suggested that the activities of steroid compounds (cholestanol, cholesterol, ketone, and a number of androstane derivatives) were active in the musk gland.

We detected a total of 8986 DEGs in the musk gland of Chinese forest musk deer 6068 of which were upregulated and 2918 downregulated, as compared with the same genes in heart tissue. We screened 17 significantly enriched GO terms that were correlated with the metabolic pathways of musk compounds. Most of the terms were annotated

in biological processes including aldosterone metabolic process, flavone metabolic process, aldosterone biosynthetic process and terpenoid biosynthetic process. We found out that although only two pathways (GnRH signalling and steroid hormone biosynthesis) were significantly enriched in DEGs ($P < 0.05$), yet all the pathways might play critical roles in musk production.

We discovered several well-studied genes of known function in the analysis, including *STS*, *UGT1A3*, *HSD17B7*, *CYP1B1* and *SRD5A1*, all annotated by GnRH signalling and steroid hormone biosynthesis. *STS* encodes a multi-pass membrane protein that is localized to the endoplasmic reticulum. It belongs to the sulphatase family, and hydrolyzes several 3-beta-hydroxysteroid sulphates, which serve as metabolic precursors for oestrogens, androgens and cholesterol (Jiang *et al*. 2016). *UGT1A3* encodes an enzyme of the glucuronidation pathway that transforms small lipophilic molecules (steroids, bilirubin, hormones and drugs) into water-soluble, excretable metabolites (Caillier *et al*. 2007). *HSD17B7* encodes an enzyme that functions both as a 17-beta-hydroxysteroid dehydrogenase in the biosynthesis of sex steroids, and as a 3-ketosteroid reductase in the biosynthesis of cholesterol (Wang *et al*. 2015). *CYP1B1* encodes cytochrome P450 proteins, which catalyze many reactions involved in drug metabolism and the synthesis of cholesterol, steroids and other lipids (Wang *et al*. 2015). *SRD5A1* catalyzes the conversion of testosterone into the more potent androgen, dihydrotestosterone (DHT) (Crowley *et al*. 2014). Further studies are needed to explore the association of these genes with musk secretion.

To our knowledge, this is the first *de novo* assembly and analysis (without a reference genome) of the Chinese forest musk deer transcriptome. This dataset represents a large number of genes that can be utilized for further exploration of musk production mechanisms, and for seeking efficient and scientific means to protect and utilize *Moschus* resources. Further studies with more samples and tissues should incorporate reverse transcription-PCR (RT-PCR) and other molecular technologies to investigate the specific roles of candidate genes in this species. The transcriptome and DEGs that we obtained will provide significant information to the scientific community and help towards expediting studies on Chinese forest musk deer and other mammals with scent glands, such as muskrat, the small Indian civet and beaver.

Acknowledgements

This study was supported by National Natural Science Foundation of China (NSFC31672396 and NSFC81503188); the programme from Sichuan Agricultural University (02920400) and Sichuan Provincial Department of Science and Technology Programme (2015JQO023).

References

Anders S. and Huber W. 2013 Differential expression of RNA-seq data at the gene level-the DESeq package (available at www.bioconductor.org/packages/devel/bioc/vignettes/DESeq/inst/doc/DESeq.pdf).

Audic S. and Claverie J. M. 1997 The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995.

Bi S. Z. and Shen Y. 1986 Study on morphology and chemical communication mechanism in musk gland. *Chinese J. Zool.* **02**, 11–14.

Bi S. Z., Shen Y., Zhu D. X., Zhu C. S., Jia L. Z. and Zhang Z. M. 1984 Study on ultra-microstructure and musk secretion of musk gland in prosperous secreting stage. *Acta Ther. Sinica* **02**, 81–85.

Caillier B., Lépine J., Tojcic J., Ménard V., Perusse L., Bélanger A. *et al*. 2007 A pharmacogenomics study of the human estrogen glucuronosyltransferase *UGT1A3*. *Pharmacogenet Gen.* **17**, 481–495.

Cao X. H. and Zhou Y. D. 2007 Progress on anti-inflammatory effects of musk. *China pharm.* **18**, 1662–1665.

Chen H. Y., Shen H., Jia B., Zhang Y. S., Wang X. H. and Zeng X. C. 2015 Differential gene expression in ovaries of Qira black sheep and Hetian sheep using RNA-Seq technique. *PLoS One* **10**, e0120170.

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. MD thesis, Zhejiang University, Hangzhou.

Conesa A., Götz S., Garcíagómez J. M., Terol J., Talón M. and Robles M. 2005 Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674.

Crowley R. K., Hughes B., Gray J., McCarthy T., Hughes S., Shackleton C. H. *et al*. 2014 Longitudinal changes in glucocorticoid metabolism are associated with later development of adverse metabolic phenotype. *Eur. J. Endocrinol*. **171**, 433–442.

Deng F. M. 1986 Domestication and grazing control of forest musk deer (*Moschus berezovskii*). *Chinese J. Wildlife* **4**, 35–37.

Deng T., Pang C., Lu X., Zhu P., Duan A. and Tan Z. 2016 *De novo* transcriptome assembly of the Chinese swamp buffalo by RNA sequencing and SSR marker discovery. *PLoS One* **11**, e0147132.

Feng Q. Q. and Liu T. J. 2015 Progress on pharmacological activity of muscone. *Food Drug* **3**, 212–214.

Grabherr M. G., Haas B. J., Moran Y., Levin J. Z., Thompson D. A. and Ido A. 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Guan T. L., Zeng B., Peng Q. K., Yue B. S. and Zou F. D. 2009 Microsatellite analysis of the genetic structure of captive forest musk deer populations and its implication for conservation. *Biochem. Syst. Ecol.* **37**, 166–173.

Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D. and Bowden J. 2013 *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc*. **8**, 1494–1512.

Harris M. A., Clark J. and Ireland A. 2004 The gene ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, 258–261.

Jia B. Y., Ba H. X., Wang G. W., Yang Y., Cui X. Z. *et al*. 2016 Transcriptome analysis of sika deer in China. *Mol. Genet. Genomics* **291**, 1–13.

Jiang M., Klein M., Zanger U. M., Mohammad M. K., Cave M. C., Gaikwad N. W. *et al*. 2016 Inflammatory regulation of steroid sulfatase: A novel mechanism to control estrogen

homeostasis and inflammation in chronic liver disease. *J. Hepatol*. **64**, 44–52.

Kanehisa M. 2008 The KEGG database. In *In silico simulation of biological processes: novartis foundation symposium* **247**. (ed. G. Bock and J. A. Gorde). John Wiley, Chichester (https://doi.org/10.1002/0470857897.ch8).

Li B. and Dewey C. N. 2011 Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 93–99.

Li D. Y., Chen B. L., Zhang L., Gaur U., Ma T. Y. and Jie H. 2016 The musk chemical composition and microbiota of Chinese forest musk deer males. *Sci. Rep.* **6**, 18975.

Liu M., Yao B., Zhang H., Guo H., Hu D. and Wang Q. 2014 Identification of novel reference genes using sika deer antler transcriptome expression data and their validation for quantitative gene expression analysis. *Genes Genom*. **36**, 573–582.

Martin M. 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J.* **17**, 10–12.

Martin J. A. and Wang Z. 2011 Next-generation transcriptome assembly. *Nat. Rev. Genet*. **12**, 671–682.

Mortazavi A., Williams B. A., Mccue K., Schaeffer L. and Wold B. 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628.

Moriya Y., Itoh M., Okuda S., Yoshizawa A. C. and Kanehisa M. 2007 Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, 182–185.

Ockendon N. F., O'Connell L. A., Bush S. J., Monzón-Sandoval J., Barnes H. and Székely T. 2015 Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Mol. Ecol. Resour*. **16**, 446–458.

Peng H., Liu S., Zou F., Zeng B. and Yue B. 2009 Genetic diversity of captive forest musk deer (*Moschus berezovskii*) inferred from the mitochondrial DNA control region. *Anim. Genet.* **40**, 65–72.

Schmieder R. and Edwards R. 2011 Quality control and preprocessing of metagenomic datasets. *Bioinforma Oxf. Engl*. **27**, 863–864.

Seth S. D., Mukhopadhyay A. B. and Prbhakar M. C. 1973 Antihista–minic and spasmolytic effects of musk. *JPN. J. Pharmacol.* **23**, 673–679.

Sheng H. L. 1996 Protection and utilization of musk deer resources in China. *Chinese J. Wildlife* **91**, 10–12.

Strickler S. R., Aureliano B. and Mueller L. A. 2012 Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Am. J. Bot*. **99**, 257–266.

Su B., Wang Y. X. and Wang Q. S. 2001 Mitochondrial DNA sequences imply Anhui musk deer a valid species in genus *Moschus*. *Zool. Res*. **22**, 169–173.

Wang X., Gérard C., Thériault J. F., Poirier D., Doillon C. J. and Lin S. X. 2015 Synergistic control of sex hormones by 17β–HSD type 7: a novel target for estrogen-dependent breast cancer. *J. Mol. Cell Biol*. **7**, 568–579.

Wang Y. and Harris R. 2015 *Moschus berezovskii*. (errata version published in 2016) The IUCN Red List of Threatened Species 2015: e.T13894A103431781. (downloaded on 09 January 2017).

Wang Z., Li M., Li L., Sun H. and Lin X. Y. 2015 Association of single nucleotide polymorphisms in the *CYP1B1* gene with the risk of primary open-angle glaucoma: a meta-analysis. *Genet. Mol. Res*. **14**, 17262–17272.

Xie C., Mao X., Huang J., Ding Y., Wu J., Dong S. *et al*. 2011 KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res*. **39**, 316–322.

Young M. D., Wakeeld M. J., Smyth G. K. and Oshlack A. 2012 goseq: Gene ontology testing for RNA-seq datasets. 1–25.

Zhang Z., Deng Z. and Li Z. C. 1985 Domestication and trans-cultivation of forest musk deer (*Moschus berezovskii*). *J. Chinese Med. Mater*. **2**, 14–15.

Zhao S. S. 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. Ph.D. thesis, Zhejiang University, Hangzhou.

Zhao Y., Yao B., Zhang M., Wang S., Zhang H. and Xiao W. 2013 Comparative analysis of differentially expressed genes in sika deer antler at different stages. *Mol. Biol. Rep.* **40**, 1665–1676.

Corresponding editor: INDRAJIT NANDA