GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

De novo assembly and characterization of skin transcriptome using RNAseq in sheep (*Ovis aries*)

Y.J. Yue[1,2], J.B. Liu[2], M. Yang[2], J.L. Han[2], T.T. Guo[2], J. Guo[2], R.L. Feng[2] and B.H. Yang[1,2]

[1]College of Animal Science and Technology, Gansu Agricultural University, Lanzhou, China
[2]Lanzhou Institute of Husbandry and Pharmaceutical Sciences, Chinese Academy of Agricultural Sciences, Lanzhou, China

Corresponding author: B.H. Yang
E-mail: yangbh2004@163.com

Genet. Mol. Res. 14 (1): 1371-1384 (2015)
Received December 18, 2013
Accepted May 8, 2014
Published February 13, 2015
DOI http://dx.doi.org/10.4238/2015.February.13.16

ABSTRACT. Wool is produced via synthetic processes of wool follicles, which are embedded in the skin of sheep. The development of new-generation sequencing and RNA sequencing provides new approaches that may elucidate the molecular regulation mechanism of wool follicle development and facilitate enhanced selection for wool traits through gene-assisted selection or targeted gene manipulation. We performed *de novo* transcriptome sequencing of skin using the Illumina Hiseq 2000 sequencing system in sheep (*Ovis aries*). Transcriptome *de novo* assembly was carried out via short-read assembly programs, including SOAPdenovo and ESTScan. The protein function, clusters of orthologous group function, gene ontology function, metabolic pathway analysis, and protein coding region prediction of unigenes were annotated by BLASTx, BLAST2GO, and ESTScan. More than 26,266,670 clean reads were collected and assembled into 79,741 unigene sequences, with a final assembly length of 35,447,962 nucleotides. A total of 22,164 unigenes were annotated, accounting for

 ©FUNPEC-RP www.funpecrp.com.br

36.27% of the total number of unigenes, which were divided into 25 classes belonging to 218 signaling pathways. Among them, there were 17 signal paths related to hair follicle development. Based on mass sequencing data of sheepskin obtained by RNA-Seq, many unigenes were identified and annotated, which provides an excellent platform for future sheep genetic and functional genomic research. The data could be used for improving wool quality and as a model for human hair follicle development or disease prevention.

Key words: Sheep (*Ovis aries*); Skin; *De novo* assembly; Transcriptome; RNA sequencing

INTRODUCTION

Wool production is a major agricultural industry throughout the world; the most important wool-producing countries are Australia, China, New Zealand, South Africa, and a number of countries in South America (Galbraith, 2010a). China, for example, as the world's second largest producer of wool, accounting for ~10% of the world's wool production (i.e., 12 million tons) (FAO, 2012). When compared to the quality of Australian wool, the wool from China has a broader micro and a shorter staple; most of the wool produced in China is not suitable for the production of high-end woolen apparel (Liu et al., 2011). The current annual wool processing capacity in China exceeds 400 kt (clean equivalent). However, China is not able to produce enough wool to meet the demand of the processors, resulting in an increase in imports. In 2009, total wool imports into China reached 327 kt, accounting for 33% of the world's output (Liu et al., 2011). Therefore, there is an urgent need to improve the quantity and quality of wool by researching the genetics, endocrinology, and ectoparasitic diseases of sheep (Smith et al., 2010).

Wool is produced via synthetic processes using wool follicles, which are embedded in the skin of sheep (Galbraith, 2010b). The underlying molecular mechanisms regulating wool follicle initiation, development, and cycling can facilitate enhanced selection for wool traits through gene-assisted selection or targeted gene manipulation (Norris et al., 2005). In sheep, wool follicle induction and morphogenesis have been characterized in a number of studies (McCloghry et al., 1992; Bond et al., 1994; Hynd et al., 1999). Gene expression profiles of sheepskin have also been generated by sequencing expressed sequence tages (ESTs) and complementary DNA (cDNA) microarray (Adelson et al., 2004; Norris et al., 2005; Smith et al., 2010; Burgess et al., 2012; Peñagaricano et al., 2012). However, molecular regulation of these developmental processes has been hampered by the paucity of information on the sheep genome and skin transcriptome (Jager et al., 2011). Most of our current and rapidly growing knowledge on the genes controlling skin and fiber development is provided by studies on the hair follicles of mice and humans (Botchkarev and Paus, 2003; Schmidt-Ullrich and Paus, 2005; Bostjancic and Glavac, 2008; Wang et al., 2012). Currently, the development of new-generation sequencing and RNA sequencing (RNA-Seq) provides new approaches that may elucidate the molecular regulation mechanism of hair follicle development (Jager et al., 2011; Okano et al., 2012; Geng et al., 2013). New-generation sequence technologies have opened the door to genome-scale experiments in organisms that lack comprehensive genome or transcriptome information, thus making it possible to assemble novel transcripts and identify differential regulation in a single experiment (Birzele et al., 2010; Sun et al., 2010; Abyzov et al., 2012).

 ©FUNPEC-RP www.funpecrp.com.br

In this study, we performed *de novo* transcriptome sequencing of skin using the Illumina Hiseq 2000 sequencing system in sheep (*Ovis aries*). More than 26,266,670 clean reads were collected and assembled into 79,741 unigenes. Annotation and gene ontology analyses were then performed on these unigenes, thus providing a valuable resource for future genetic and genomic research on sheep and closely related species

MATERIAL AND METHODS

Animal material

Gansu Alpine fine wool sheep were bred in the Huang Cheng District of Gansu Province, China, by cross breeding Mongolian or Tibetan sheep with Xinjiang Fine Wool sheep and then with some fine wool sheep breeds from the Union of Soviet Socialist Republics, such as Caucasian and Salsk. The breed was approved by the Gansu provincial government in 1980. Since then, to improve wool quality, the breed was hybridized with Australian Merino, New Zealand Merino, and German Mutton Merino sheep. Today, the modern Gansu Alpine fine wool sheep can be classified into 3 strains according to the fineness of the wool, including superfine [fiber diameter (FD) < 19.0 μm], fine (19.0 μm < FD < 21.0 μm) and dual-purpose strains (FD > 21.0 μm). A body side skin sample (SHEawmTARAAPEI-12) was collected from the superfine strain (FD = 16.5 μm) of Gansu Alpine fine wool sheep. The sampled tissues were immediately frozen in liquid nitrogen and stored at -80°C for subsequent analysis.

RNA extraction

Total RNA was isolated from the tissues using the RNeasy Maxi Kit (Qiagen, Hilden, Germany) according to manufacturer instructions. RNA quality was verified using a 2100 Bioanalyzer RNA Nanochip (Agilent, Santa Clara, CA, USA), and the RNA Integrity Number value was >8.5. Then, the RNA was quantified using the Nano Drop ND-2000 Spectrophotometer (Nano-Drop, Wilmington, DE, USA).

cDNA library construction and sequencing

Illumina sequencing using the GAII platform was performed at the Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China (http://www.genomics.cn/index.php) according to manufacturer instructions (Illumina, San Diego, CA, USA). Beads with Oligo (dT) were used to isolate poly(A) messenger RNA (mRNA) after total RNA was collected from eukaryote (prokaryocyte can be treated with a kit to remove ribosomal RNA before the next step). Fragmentation buffer was added for interrupting mRNA to short fragments. Utilizing these short fragments as templates, a random hexamer-primer was used to synthesize the first-strand cDNA. The second-strand cDNA was synthesized using a buffer, dNTPs, RNaseH, and DNA polymerase I. Short fragments were purified with a QiaQuick polymerase chain reaction (PCR) extraction kit and resolved with EB buffer for end reparation and adding the poly(A). Subsequently, the short fragments were connected with sequencing adapters. Then, after agarose gel electrophoresis, the suitable fragments were selected for PCR amplification as templates. Finally, the library was sequenced using Illumina HiSeq™ 2000.

Data filtering and *de novo* assembly

The quality requirement for *de novo* transcriptome sequencing was far higher than that for resequencing because sequencing errors can cause difficulties for the short-read assembly algorithm. Therefore, we carried out a stringent filtering process. First, we removed reads that did not pass the built-in Illumina's software failed-chastity filter according to the relation "failed-chastity ≤ 1" using a chastity threshold of 0.6 on the first 25 cycles. Second, we discarded all reads with adaptor contamination. Third, we ruled out low-quality reads with ambiguous sequences "N". Finally, the reads with more than 10% Q < 20 bases were also removed.

De novo assembly

Transcriptome *de novo* assembly was carried out with the short-read assembly program SOAPdenovo (Li et al., 2010). SOAPdenovo first combined reads with a certain length of overlap to form longer fragments without N, which are called contigs. Then, the reads were mapped back to contigs; with paired-end reads, it is able to detect contigs from the same transcript as well as the distances between those contigs. Next, SOAPdenovo connected the contigs using N to represent unknown sequences between 2 contigs and scaffolds were formed. Paired-end reads were used again for gap filling of scaffolds to form sequences with the least number of Ns that could not be extended on either end (i.e., unigenes). When multiple samples from the same species are sequenced, unigenes from the assembly of each sample can be further processed for sequence splicing and redundancy removal by using the sequence clustering software to acquire the longest possible non-redundant (nr) unigenes. In the final step, BLASTx alignment (E value < 0.00001) between unigenes and protein databases [e.g., nr database, Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and clusters of orthologous groups (COG)] was performed, and the best-fit alignments were used to determine the sequence direction of unigenes. If the results obtained from different databases conflicted with each other, a priority order of nr, Swiss-Prot, KEGG, and COG was followed to determine the sequence direction of unigenes. When a unigene did not align to any of the above databases, the ESTScan software (Iseli et al., 1999) was introduced to predict the coding regions and determine the direction of the sequence. For unigenes with sequence directions, we provided their sequences from the 5'- to 3'-end; for those without a direction, we provided their sequences from the assembly software.

Unigene COG function annotation

Unigene annotation provides information on the expression and functional annotation of a unigene. Unigene sequences were first aligned by BLASTx to protein databases (e.g., nr, Swiss-Prot, KEGG, and COG; E value < 0.00001), retrieving proteins with the highest sequence similarity with the given unigenes along with their protein functional annotations. COG is a database where orthologous gene products are classified. Every protein in the COG is assumed to have evolved from an ancestral protein, and the whole database is built on coding proteins with complete genomes and evolutionary relationships of bacteria, algae, and

eukaryotic creatures. Unigenes were aligned to the COG database to predict and classify possible functions of unigenes.

Unigene Gene Ontology (GO) classification

We can obtain GO functional annotation with nr annotation. GO is an international standardized gene functional classification system that offers a dynamically updated and controlled vocabulary; it is a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has 3 ontologies, including molecular function, cellular component, and biological process. The basic unit of GO is the GO-term. Every GO-term belongs to a type of ontology.

With nr annotation, we used the BLAST2GO program (Conesa et al., 2005) to obtain GO annotation of unigenes. BLAST2GO has been cited by other articles (i.e., >150 times) and is a widely recognized GO annotation software. After obtaining the GO annotation for each unigene, we used the WEGO software (Ye et al., 2006) to conduct GO functional classification for all unigenes to investigate the species distributions of gene functions at the macro level.

Unigene metabolic pathway analysis

KEGG is a database that is able to analyze a gene product during metabolic processes and related gene function in cellular processes. KEGG may further research on the genetics of biologically complex behaviors. Using KEGG annotation, we can comment on the pathways for unigenes.

Protein coding DNA sequence (CDS) prediction

Unigenes were first aligned by BLASTx (E value < 0.00001) with protein databases in the priority order of nr, Swiss-Prot, KEGG, and COG. Unigenes aligned with databases with higher priority did not enter the next circle. The alignments ended when all of the circles were completed. Proteins with the highest ranks in the BLAST results were utilized to determine the CDSs of unigenes; then, the coding region sequences were translated into amino sequences with the standard codon table. Thus, both the nucleotide sequences (5'-3') and amino sequences of the unigene coding regions were acquired. Unigenes that could not be aligned with any database were scanned by ESTScan (Iseli et al., 1999) to obtain nucleotide (5'-3') and amino sequences of the coding regions.

RESULTS

Sheepskin transcriptome sequencing and assembly

The amount of sequencing data is an important indicator for completion of transcriptome sequencing. The project extracted tissue mRNA from sheepskin and then conducted the transcriptome sequencing. We obtained 26,266,670 clean reads, and the number of bases obtained by the clean reads was 2,364,000,300 nucleotides (nt) (Table 1). Transcriptome *de novo* assembly was carried out with the short-read assembly program SOAPdenovo using

mRNAs, ESTs, and clean reads following a bioinformatic workflow (Li et al., 2010; Jager et al., 2011; NCBI, 2013). A total of 40,300 ovine gene GenBank entries (corresponding to 17,319 unigenes) and 370,194 ovine ESTs were available for assembly. SOAPdenovo output of 853,437 contigs with a final assembly length of 105,879,352 nt. The average size of the contigs was 134 nt. The shortest assembly contig was 50 bp, and the longest contig was 4881 bp; there were 779,172 contigs that were <200 bp in length, which accounted for 91.30% of the contigs (Table 2).

Table 1. Output statistics of sequencing.

Sample	Total reads	Total nucleotides (nt)	Q20 (%)	N (%)	GC (%)
SHEawmTARAAPEI-12	26,266,670	2,364,000,300	92.92%	0.00%	48.93%

Q20 = quality score of 20; Total nucleotides = total reads 1 x read 1 size + total reads 2 x read 2 size.

Table 2. Contig quality of sheepskin.

Length of contig	Number	Percent
75-100 nt	685,968	80.38%
100-200 nt	93,204	10.92%
200-300 nt	33,428	3.92%
300-400 nt	15,112	1.77%
400-500 nt	8,197	0.96%
≥500 nt	17,528	2.05%
N50	90	
Mean	124	
All contig	853,437	
Length of all contig (nt)	105,879,352	

To determine different contigs from the same transcript and the distance between these contigs by paired-end reads, we used the sequence-obtained clean reads for alignment with the contigs. Next, SOAPdenovo connected the contigs using N to represent unknown sequences between 2 contigs; then, scaffolds were constructed. In this study, we obtained 147,155 scaffolds with a final assembly length of 44,758,530 nt. The average size of the scaffolds was 304 nt (Table 3). The length distribution of the assembled scaffolds primarily included small fragments. The number of scaffolds that were <500 bp in length totaled 126,021 (i.e., 85.64%), the proportion of which was significantly reduced. The number of scaffolds that did not contain nucleotide deletions was 122,860 (i.e., 83.49%; **Figure S1**).

Table 3. Scaffold quality of sheepskin.

Length of scaffold	Number	Percent
100-500 nt	126,021	85.64%
500-1000 nt	16,004	10.88%
1000-1500 nt	3,775	2.57%
1500-2000 nt	1,038	0.71%
≥2000 nt	317	0.22%
N50	387	
Mean	304	
All scafSeq	147,155	
Length of all scafSeq (nt)	44,758,530	

Paired-end reads were used again for gap filling of scaffolds to obtain sequences with the least Ns and that could not be extended on either end (i.e., unigenes). A total of 79,741 sheepskin unigene sequences with a final assembly length of 35,447,962 nt were obtained, which contained the least Ns and could not be extended on either end. The average sequence length was 445 bp. A total of 74,592 unigenes were <1000 bp in length (i.e., 93.55%). The N50 length of 508 bp (i.e., half of the assembled bases were incorporated into unigenes with a length at least 508 bp) was obtained; 26.53% (21,151 unigenes) had lengths >500 bp (Table 4). A total of 84.67% of the unigenes had no gaps, and <0.54% of the unigenes had a gap percentage (ratio of number 'N' to unigene length) >20% (**Figure S2**). Only one unigene had a base deletion ratio of >0.3. Analysis of the positional distribution of obtained sequences identified several characteristics as follows: 1) sheepskin transcriptome sequencing obtained a relatively large number of reads and a more balanced distribution of the unigene 5'-end; and 2) the number of reads located in the unigene 3'-end was less than that in the 5'-end. For the number of reads with a relative increase in position, there was a linear downward trend in the relative position of 5' to 3' that was >0.6 (Figure 1). This result is consistent with other studies (Wang et al., 2010; Shi et al., 2011). This indicates that the transcriptome sequencing quality of the sheepskin and that of other non-model organisms was considerable.

Table 4. Unigene quality of sheepskin.

Length of unigene	Number	Percent
100-500 nt	58,590	73.48%
500-1000 nt	16,002	20.07%
1000-1500 nt	3795	4.76%
1500-2000 nt	1,036	1.30%
≥2000 nt	318	0.40%
N50	508	
Mean	445	
All unigenes	79,741	
Length of all unigenes (nt)	35,447,962	

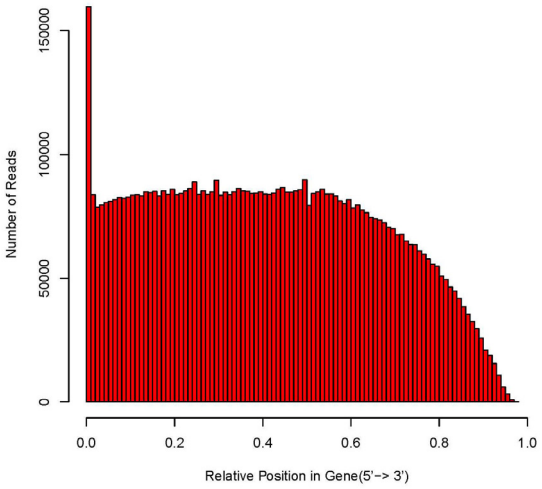

Figure 1. Randomness of reads from sample.

Functional annotation of sheepskin transcriptome

We used the BLASTx comparison to align the unigene sequences with the protein databases nr, Swiss-Prot, KEGG, and COG (E value < 0.00001). We searched for the highest protein sequence similarity for a given unigene and gained functional annotation information on sheepskin unigene proteins. The functional annotation results of sheepskin transcriptome sequencing are shown in Table 1. After alignment with the 4 databases, a total of 22,164 sheepskin unigenes were annotated, accounting for 36.27% of the total number of unigenes (i.e., a total of 79,741). Of these, 28,924 were annotated in the nr database; 26,079 in the Swiss-Prot database; 17,113 in the KEGG database; and 6616 in the COG database. A total of 6616 unigenes were annotated in all 4 databases.

Through BLAST and the National Center for Biotechnology Information (NCBI) COG databases, we obtained 12,711 sheepskin unigenes via COG functional classification. These were divided into 25 categories. The number of unigenes belonging to the R class (general function prediction only) was 2083 (i.e., 16.39%). The J class (chromatin structure and dynamics) contained 1253 unigenes (i.e., 9.86%). The L class (energy production and conversion) contained 1147 unigenes (i.e., 9.02%). Only one unigene fell into the W class (extracellular structures) (Figure 2).

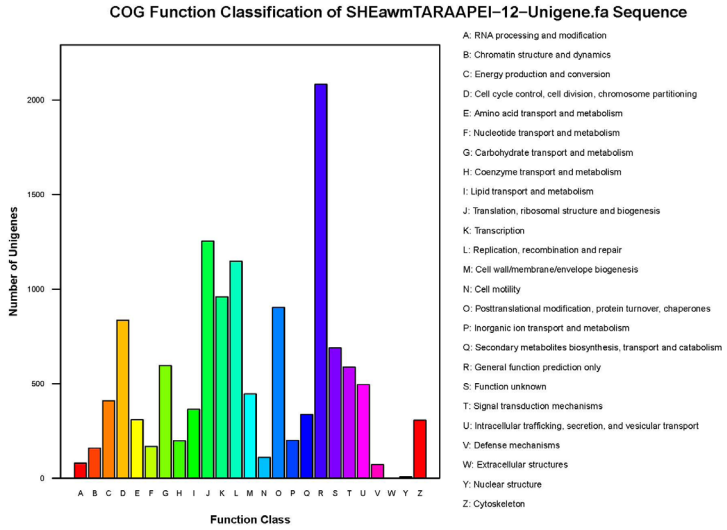

Figure 2. COG functional classification of the SHEawmTARAAPEI-12-Unigene.

More than 11,726 unigenes were classed into 3 GO categories, including cellular component, biological process, and molecular function, which were related to 53 types of biological functions. Under the cellular component category, large numbers of unigenes were categorized as cell (9669 unigenes), cell part (9669 unigenes), organelle (6731 unigenes), and organelle part (3490 unigenes). For the biological process category, cellular process (8199 unigenes), biological regulation (4371 unigenes), and pigmentation (4094 unigenes) represented the greatest proportion of unigenes. Under the molecular function category, binding (8799 unigenes) and catalytic activity (4200 unigenes) were the top 2 most abundant subcategories (Figure 3).

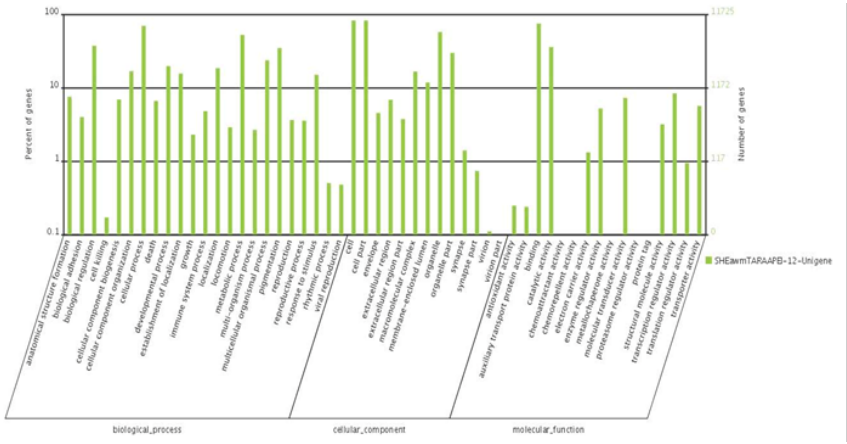

Figure 3. GO functional classification of the SHEawmTARAAPEI-12-Unigene.

There were 17,096 unigenes mapped into 218 KEGG pathways. The pathways or maps with the highest representation were metabolic pathways (1977 unigenes, 11.56%, ko01100), focal adhesion (966 unigenes, 5.65%, ko04510), and regulation of the actin cytoskeleton (840 unigenes, 4.91%, ko04810). Signaling pathways related to hair follicle development included, for example, the hedgehog signaling, insulin signaling, Jak-STAT signaling, MAPK signaling, melanogenesis, notch signaling, TGF-beta signaling, Toll-like receptor signaling, VEGF signaling, and Wnt signaling pathways (Botchkarev and Paus, 2003; Adelson et al., 2004; Norris et al., 2005; Schmidt-Ullrich and Paus, 2005; Bostjancic and Glavac, 2008; Smith et al., 2010; Penagaricano et al., 2012; Wang et al., 2012) (Table 5).

Table 5. Signaling pathways related to hair follicle development in sheepskin.

Pathway	DEGs with pathway annotation (17,096 unigenes)	Pathway ID
Hedgehog signaling pathway	83 (0.49%)	ko04340
Insulin signaling pathway	317 (1.85%)	ko04910
Jak-STAT signaling pathway	178 (1.04%)	ko04630
Lysine biosynthesis	15 (0.09%)	ko00300
Lysine degradation	259 (1.51%)	ko00310
MAPK signaling pathway	536 (3.14%)	ko04010
Melanogenesis	171 (1%)	ko04916
Melanoma	115 (0.67%)	ko05218
Metabolic pathway	1977 (11.56%)	ko01100
Notch signaling pathway	164 (0.96%)	ko04330
PPAR signaling pathway	153 (0.89%)	ko03320
TGF-beta signaling pathway	197 (1.15%)	ko04350
Toll-like receptor signaling pathway	157 (0.92%)	ko04620
VEGF signaling pathway	157 (0.92%)	ko04370
Wnt signaling pathway	384 (2.25%)	ko04310

DEGs = differentially expressed genes; Jak-STAT = Janus kinase/signal transducers and activators of transcription; MAPK = mitogen-activated protein kinase; PPAR = peroxisome proliferator-activated receptor; VEGF = vascular endothelial growth factor; Wnt = wingless.

A total of 22,406 CDSs were obtained by BLASTx (E value < 0.00001) to protein databases in the priority order of nr, Swiss-Prot, KEGG, and COG; >3478 CDSs were obtained by

ESTScan (Table 6). A total of 21,309 and 3337 CDSs of <1000 nt in length by BLASTx and ESTScan accounted for 95.10 and 95.95%, respectively. The number of proteins with lengths <1000 nt by BLASTx and ESTScan was 22,406 and 3478 CDs, accounting for 99.99 and 100%, respectively.

Table 6. Length of coding DNA sequence (CDS) and protein of sheepskin by blastx and ESTScan.

Range of CDS/protein length	Number of CDS sequences in specified range		Number of protein sequences in specified range	
	BLASTx	ESTScan	BLASTx	ESTScan
200	4306	192	18594	2832
300	7091	1096	2330	435
400	3774	567	885	146
500	1983	609	353	47
600	1440	367	142	12
700	1098	164	58	4
800	695	164	32	1
900	537	108	7	1
1000	385	70	4	0
1100	279	47	0	0
1200	221	29	0	0
1300	159	16	0	0
1400	118	17	0	0
1500	76	14	0	0
1600	71	8	0	0
1700	37	1	1	0
1800	34	3	0	0
1900	28	2	0	0
2000	17	0	0	0
2100	13	2	0	0
2200	13	1	0	0
2300	7	0	0	0
2400	12	0	0	0
2500	4	0	0	0
2600	1	1	0	0
2700	2	0	0	0
2800	3	0	0	0
2900	1	0	0	0
3000	0	0	0	0
>3000	1	0	0	0

DISCUSSION

Currently, methods used for the acquisition and analysis of transcriptome data are mainly based on serial analysis of gene expression, hybridization chips (gene chip or microarray), and RNA-Seq technology. Gene expression profiles of sheepskin have also been generated by sequencing of ESTs and cDNA microarray (Adelson et al., 2004; Norris et al., 2005; Smith et al., 2010; Burgess et al., 2012). However, hybridization-based microarray technology is restricted to known sequences, which are unable to detect new transcripts. Meanwhile, it is also difficult to detect fusion gene transcripts, polycistronic transcription, and abnormal transcripts, thus limiting its use (Okoniewski and Miller, 2006). Compared to microarray technology, RNA-Seq technology has many unique advantages (Nagalakshmi et al., 2008). It is highly accurate and has a very low detection limitation, making it useful for a wide range of applications (Wang et al., 2009). These applications include the detection of new transcripts (unknown transcripts and rare transcripts) (Denoeud et al., 2008), functional re-

search of non-coding regions (e.g., non-coding RNA research, and small RNA precursor studies), transcript structure (e.g., UTR identification, intron boundaries identification, alternative splicing, and start codon identification), gene transcription, and detection of SNPs of a coding sequence (Cloonan et al., 2008). These applications make RNA-Seq technology a powerful tool for in-depth research on complex transcriptomes. The International Sheep Genome Consortium utilizes the results from the Human Genome Project; moreover, genomic research for the dog and cow is used to derive a virtual sheep genome map (Dalrymple et al., 2007; Archibald et al., 2010). However, these are currently available only for low-coverage genome sequences (Archibald et al., 2010). In order to reveal the molecular mechanism of hair follicle development, we performed *de novo* transcriptome sequencing of skin using the Illumina Hiseq 2000 sequencing system in sheep (*Ovis aries*). We obtained 79,741 unigenes in sheepskin, totaling 35,447,962 bp. The average sequence length was 445 bp. The quality of the sequencing results was equivalent to traditional Sanger sequencing, but the collection of the information was time-consuming when compared to that of traditional Sanger sequencing and microarray (Lobo et al., 2009). Using bioinformatic tools and aligning sequences with the Uniprot, NCBI's nr, COG, Pfam, InterPro, and KEGG 6 databases, 22,164 of the sheepskin unigenes were annotated, which only accounted for 36.27% of the total number of unigenes (total number = 79,741). Unannotated unigenes compared using ESTScan yielded 3478 CDSs. A total of 35,171 unigenes remained unannotated; therefore, additional research on the assembly and annotation of the sheep transcriptome is needed for the sheep reference genome (Archibald et al., 2010).

The skin is composed of the epidermis, dermis, and subcutaneous tissue that interconnect anatomically (Prost-Squarcioni et al., 2008). Hair follicles develop as a result of epithelial-mesenchymal interactions between epidermal keratinocytes committed to hair-specific differentiation and the clustering of dermal fibroblasts that form follicular papilla. The development of primary, secondary, and derived follicles in the pelage of Merino sheep has been described by Rogers (2006). Primary follicles with arrector pilae muscle, and sebaceous and apocrine glands commence development first (E70) in the form of a central follicle surrounded by 2 lateral primaries. Then, the secondary follicles (So) associated with the primaries appear (E85). Additional follicles develop by branching from the So follicles and are apparent by E105. Branching of the secondary original follicles is critical to the volume and nature of the Merino fleece, as secondary follicles represent the major source of fine fibers (Adelson et al., 2004; Rogers, 2006). During postnatal life, hair follicles show patterns of cyclic activity with periods of active growth and hair production (anagen), apoptosis-driven involution (catagen), and relative resting (telogen) (Botchkarev and Paus, 2003). However, Merino wool follicles are predominant throughout growth, different to many animals such as the mouse, rabbit, and guinea-pig (Rogers, 2006). During the past decade there has been a dramatic increase in our knowledge on the molecular control of fiber initiation and development in mice and humans (Botchkarev and Paus, 2003; Schmidt-Ullrich and Paus, 2005; Bostjancic and Glavac, 2008; Wang et al., 2012; Cadau et al., 2013). Gene expression profiles of sheepskin have also been generated by the sequencing of ESTs and cDNA microarray (Adelson et al., 2004; Norris et al., 2005; Smith et al., 2010; Burgess et al., 2012). However, molecular regulation of these developmental processes has been hampered by the paucity of information on the sheep genome and skin transcriptome (Jager et al., 2011) However, comparative analyses of gene expression patterns in skin suggest that the molecular regulation of the basic mechanisms of follicle initiation and development and fiber production are similar in most mammals. Studies in these mammals have detailed genes with translated products that control hair follicle ini-

 ©FUNPEC-RP www.funpecrp.com.br

tiation, development, and regulation of secondary branching events, as described in reviews (Galbraith, 2010b). Such factors defined in the hair follicle include members of the wingless, bone morphogenetic protein, fibroblast growth factor, tumor necrosis factor, transforming growth factor, and the notch and hedgehog signaling pathways. In this study, these regulatory pathways and signaling molecules are also found in the sheep skin. This has only been used for improving wool quality, but can also be used as a model for human hair follicle development or disease prevention (Archibald et al., 2010).

The most important function of the skin is to form an effective barrier between the 'inside' and 'outside' of an organism. The epidermis comprises the physical, chemical/biochemical (antimicrobial and innate immunity), and adaptive immunological barriers because it is composed of the stratum corneum, nucleated epidermis, lipids, acids, hydrolytic enzymes, antimicrobial peptides and macrophages, and humoral and cellular constituents of the immune system (Proksch et al., 2008). Therefore, in this study, we aligned sequences using the Uniprot, NCBI nr, COG, Pfam, InterPro, and KEGG 6 databases. A total of 22,164 unigenes in the sheepskin were annotated and divided into 25 categories, which included 218 signaling pathways involved in cell composition (23,822 unigenes), biological processes (49,340 unigenes), and molecular function (16,150 unigenes). Among the signaling pathways that are related to diseases, the following number of unigenes and respective signaling pathways were identified: bacteria invasion of epithelial cells: 334; vibrio cholerae infection: 219; epithelial cell signaling in *Helicobacter pylori* infection: 132; pathogenic *Escherichia coli* infection: 379; shigellosis: 362; leishmaniasis: 131; Chagas disease: 173; malaria: 71; and ameobiasis: 661. This study will also provide a foundation for the research and development of livestock breeding and antiectoparasitic drugs (Burgess et al., 2012; Losson, 2012).

CONCLUSION

Based on mass sequencing data of sheepskin obtained by RNA-Seq, many unigenes were identified and annotated, which provides an excellent platform for future genetic and functional genomic research.

ACKNOWLEDGMENTS

Research supported by the Central Level, Scientific Research Institutes for Basic R&D Special Fund Business (#BRF100102), the Earmarked Fund for Modern China Wool & Cashmere Technology Research System (#nycytx-40-3), the Project of the National Natural Science Foundation for Young Scholars of China (#31402057), and the National High Technology Research and Development Program of China ("863" Program; #2008AA101011-2).

Supplementary material

REFERENCES

Abyzov A, Mariani J, Palejev D, Zhang Y, et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* 492: 438-442.

Adelson DL, Cam GR, DeSilva U and Franklin IR (2004). Gene expression in sheep skin and wool (hair). *Genomics* 83: 95-105.

Archibald AL, Cockett NE, Dalrymple BP, Faraut T, et al. (2010). The sheep genome reference sequence: a work in progress. *Anim. Genet.* 41: 449-453.

Birzele F, Schaub J, Rust W, Clemens C, et al. (2010). Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Res.* 38: 3999-4010.

Bond JJ, Wynn PC, Brown GN and Moore GP (1994). Growth of wool follicles in culture. *In Vitro Cell Dev. Biol. Anim.* 30A: 90-98.

Bostjancic E and Glavac D (2008). Importance of microRNAs in skin morphogenesis and diseases. *Acta Dermatovenerol. Alp. Pannonica Adriat.* 17: 95-102.

Botchkarev VA and Paus R (2003). Molecular biology of hair morphogenesis: development and cycling. *J. Exp. Zool. B Mol. Dev. Evol.* 298: 164-180.

Burgess ST, Greer A, Frew D, Wells B, et al. (2012). Transcriptomic analysis of circulating leukocytes reveals novel aspects of the host systemic inflammatory response to sheep scab mites. *PLoS One* 7: e42778.

Cadau S, Rosignoli C, Rhetore S, Voegel JJ, et al. (2013). Early stages of hair follicle development: a step by step microarray identity. *Eur. J. Dermatol.* 23: 4-10.

Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5: 613-619.

Conesa A, Götz S, García-Gómez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

Dalrymple BP, Kirkness EF, Nefedov M, McWilliam S, et al. (2007). Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. *Genome Biol.* 8: R152.

Denoeud F, Aury JM, Da Silva C, Noel B, et al. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* 9: R175.

Galbraith H (2010a). Foreword. Animal fibre: connecting science and production. *Animal* 4: 1447-1450.

Galbraith H (2010b). Fundamental hair follicle biology and fine fibre production in animals. *Animal* 4: 1490-1509.

Geng R, Yuan C and Chen Y (2013). Exploring differentially expressed genes by RNA-Seq in cashmere goat (*Capra hircus*) skin during hair follicle development and cycling. *PLoS One* 8: e62704.

Hynd PI, Penno NM and Bates EJ (1999). Follicle morphogenesis *in vitro*. *Exp. Dermatol.* 8: 350-351.

Iseli C, Jongeneel CV and Bucher P (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 138-148.

Jager M, Ott CE, Grünhagen J, Hecht J, et al. (2011). Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing. *BMC Genomics* 12: 158.

Li R, Zhu H, Ruan J, Qian W, et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20: 265-272.

Liu H, Zhou ZY and Malcolm B (2011). China's wool import demand: implications for Australia. *Australasian Agribusiness Rev.* 19: 16-34.

Lobo AM, Lobo RN, Paiva SR, de Oliveira SM, et al. (2009). Genetic parameters for growth, reproductive and maternal traits in a multibreed meat sheep population. *Genet. Mol. Biol.* 32: 761-770.

Losson BJ (2012). Sheep psoroptic mange: an update. *Vet. Parasitol.* 189: 39-43.

McCloghry E, Foldes A, Hollis D, Rintoul A, et al. (1992). Effects of pinealectomy on wool growth and wool follicle density in merino sheep. *J. Pineal Res.* 13: 139-144.

Nagalakshmi U, Wang Z, Waern K, Shou C, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349.

NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41: D8-D20.

Norris BJ, Bower NI, Smith WJM and Cam GR (2005). Gene expression profiling of ovine skin and wool follicle development using a combined ovine-bovine skin cDNA microarray. *Aust. J. Exp. Agr.* 45: 867-877.

Okano J, Levy C, Lichti U, Sun HW, et al. (2012). Cutaneous retinoic acid levels determine hair follicle development and downgrowth. *J. Biol. Chem.* 287: 39304-39315.

Okoniewski MJ and Miller CJ (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* 7: 276.

Peñagaricano F, Zorrilla P, Naya H, Robello C, et al. (2012). Gene expression analysis identifies new candidate genes associated with the development of black skin spots in Corriedale sheep. *J. Appl. Genet.* 53: 99-106.

Proksch E, Brandner JM and Jensen JM (2008). The skin: an indispensable barrier. *Exp. Dermatol.* 17: 1063-1072.

Prost-Squarcioni C, Fraitag S, Heller M and Boehm N (2008). Functional histology of dermis. *Ann. Dermatol. Venereol.* 135: 1S5-20.

Rogers GE (2006). Biology of the wool follicle: an excursion into a unique tissue interaction system waiting to be re-

discovered. *Exp. Dermatol.* 15: 931-949.
Schmidt-Ullrich R and Paus R (2005). Molecular principles of hair follicle induction and morphogenesis. *Bioessays* 27: 247-261.
Shi CY, Yang H, Wei CL, Yu O, et al. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12: 131.
Smith WJ, Li Y, Ingham A, Collis E, et al. (2010). A genomics-informed, SNP association study reveals FBLN1 and FABP4 as contributing to resistance to fleece rot in Australian Merino sheep. *BMC Vet. Res.* 6: 27.
Sun C, Li Y, Wu Q, Luo H, et al. (2010). *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* 11: 262.
Wang X, Tredget EE and Wu Y (2012). Dynamic signals for hair follicle development and regeneration. *Stem Cells Dev.* 21: 7-18.
Wang XW, Luan JB, Li JM, Bao YY, et al. (2010). *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* 11: 400.
Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.
Ye J, Fang L, Zheng H, Zhang Y, et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34: W293-W297.