SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

September 2014 Vol.57 No.9: 925–935
doi: 10.1007/s11427-014-4678-2

RNA-Seq analysis of yak ovary: improving yak gene structure information and mining reproduction-related genes

LAN DaoLiang[1], XIONG XianRong[2], WEI YanLi[3], XU Tong[3], ZHONG JinCheng[1], ZHI XiangDong[2], WANG Yong[1] & LI Jian[1*]

[1]*Institute of Qinghai-Tibetan Plateau, Southwest University for Nationalities, Chengdu 610041, China;*
[2]*College of Life Science and Technology, Southwest University for Nationalities, Chengdu 610041, China;*
[3]*Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China*

Received August 15, 2013; accepted January 1, 2014; published online June 6, 2014

RNA-Seq, a high-throughput (HT) sequencing technique, has been used effectively in large-scale transcriptomic studies, and is particularly useful for improving gene structure information and mining of new genes. In this study, RNA-Seq HT technology was employed to analyze the transcriptome of yak ovary. After Illumina-Solexa deep sequencing, 26826516 clean reads with a total of 4828772880 bp were obtained from the ovary library. Alignment analysis showed that 16992 yak genes mapped to the yak genome and 3734 of these genes were involved in alternative splicing. Gene structure refinement analysis showed that 7340 genes that were annotated in the yak genome could be extended at the 5′ or 3′ ends based on the alignments been the transcripts and the genome sequence. Novel transcript prediction analysis identified 6321 new transcripts with lengths ranging from 180 to 14884 bp, and 2267 of them were predicted to code proteins. BLAST analysis of the new transcripts showed that 1200–4933 mapped to the non-redundant (nr), nucleotide (nt) and/or SwissProt sequence databases. Comparative statistical analysis of the new mapped transcripts showed that the majority of them were similar to genes in *Bos taurus* (41.4%), *Bos grunniens mutus* (33.0%), *Ovis aries* (6.3%), Homo *sapiens* (2.8%), *Mus musculus* (1.6%) and other species. Functional analysis showed that these expressed genes were involved in various Gene Ontology (GO) categories and Kyoto Encyclopedia of Genes and Genomes pathways. GO analysis of the new transcripts found that the largest proportion of them was associated with reproduction. The results of this study will provide a basis for describing the normal transcriptome map of yak ovary and for future studies on yak breeding performance. Moreover, the results confirmed that RNA-Seq HT technology is highly advantageous in improving gene structure information and mining of new genes, as well as in providing valuable data to expand the yak genome information.

yak, ovary, transcriptome, RNA-Seq, improvement of gene structure, reproduction

Citation: Lan DL, Xiong XR, Wei YL, Xu T, Zhong JC, Zhi XD, Wang Y, Li J. RNA-Seq analysis of yak ovary: improving yak gene structure information and mining reproduction-related genes. Sci China Life Sci, 2014, 57: 925−935, doi: 10.1007/s11427-014-4678-2

A transcriptome is the sum of all of the RNA products, including protein-coding mRNA and non-coding RNA, of a specific tissue or cell in a particular environment or under particular physiological conditions. The transcriptome connects the genome and proteome information [1,2]. Transcriptomic analysis is important in the post-genomic era because it allows gene expression and regulation to be studied; such analysis leads to the discovery of functional genes, which serve as a starting point for the study of gene function and structure [3]. Massive-scale transcriptome detec-

*Corresponding author (email: lijian@swun.edu.cn)

tion technologies have moved from the initial methods, such as differential hybridization, mRNA differential display, and suppression subtractive hybridization, to current methods, such as DNA microarray and serial analysis of gene expression (SAGE). With the recent developments in high-throughput (HT) sequencing technologies, transcriptome sequencing (RNA sequencing or RNA-Seq) has become a more effective method for large-scale transcriptomic studies, thereby subverting the traditional transcriptomic methods [4].

RNA-Seq can rapidly obtain the sequences of almost all complementary (cDNA) transcripts from the RNA of a specific organ or tissue under certain conditions at the single-nucleotide level [5]. RNA-Seq is advantageous over other traditional transcriptomic technologies because of its HT, low cost, high sensitivity, and ability to detect low-abundance genes. Compared with DNA microarray technology, RNA-Seq does not require probes that are specific to known genes to be designed; therefore, RNA-Seq can analyze the whole transcripts of species for which no genomic data are available. Moreover, RNA-Seq is not limited by cross-reactivity and background noise problems caused by fluorescence analog signals in microarray hybridization, which significantly improves the resolution [5,6]. RNA-Seq has been used extensively in studies on gene transcription, structural variations in transcripts (e.g., alternative splicing), functions of non-coding RNA, and in the development of single nucleotide polymorphism (SNP) or simple sequence repeat (SSR) markers. The improvements in gene structure information and mining of new genes through RNA-Seq have attracted increasing attention in recent years [7–10]. The structure of genes in a genome, including the identification of 5′/3′ boundaries and untranslated regions (UTRs), can be more accurately determined using the genomic distribution and pairing information of reads obtained by RNA-Seq. Given that the annotation of transcripts in the existing databases may not be comprehensive, new transcripts and genes can be discovered by aligning the reads to known genes in the sequence databases and analyzing the annotation of the existing genes.

Yak (*Bos grunniens*), also known as the "plateau ship," is a species that is distributed mainly in the Qinghai-Tibetan Plateau and adjacent alpine or subalpine regions in China. Yaks can adapt well to the alpine grassland environment, but they can also live freely and reproduce under the harsh plateau environmental conditions, such as thin air, cold temperatures, and short grass. Yaks provide milk, meat, wool, service force, and fuel for local pastoralists. Yaks are necessary and important breeds in the local region. In addition, yaks are important contributors to the gene pool. However, compared with ordinary cattle that live in the plains, yaks reach sexual maturity more slowly and generally have lower fertility. The recent completion of the yak genomic sequence [11] provides an actionable "blueprint" for understanding various aspects of yaks, such as molecular breeding, environmental adaptation, and reproductive performance. Most of the genes in the yak genome were annotated based on bioinformatics predictions; therefore, some gene annotations may be incomplete or missing. Yak transcriptome-related studies are still limited. Therefore, in the present study, RNA-Seq technology was used to analyze and describe the normal transcriptome map of yak ovary. The results of this study will serve as a basis for future studies on yak breeding performance and will provide valuable data for the improvement of yak gene structure information and for mining for novel candidate genes.

1 Materials and methods

1.1 Sample collection

Three adult (four-year-old) healthy female Maiwa yaks with similar body sizes were selected randomly from a plateau slaughterhouse in Hongyuan, Sichuan, China (31°51′N to 33°19′N and 101°51′E to 103°23′E; average altitude of 3600 m above sea level). The ovarian tissues were collected immediately after slaughter and frozen in liquid nitrogen for RNA extraction. This study was approved by the ethics committee of Southwest University for Nationalities (Chengdu, China). Approval from the animal use and care committee was not required for this study because the samples were obtained from government-inspected slaughter facilities.

1.2 RNA extraction

Total RNA of the three yak ovarian tissues was extracted using TRIzol reagent (Life Technologies, Carlsbad, CA, USA), in accordance with the manufacturer's instructions. To remove residual genomic DNA, each RNA sample was incubated with 10 units of DNA-free DNAse I (TaKaRa (Dalian), Japan) for 30 min at 37°C. After the total RNA concentration of the three RNA samples was determined, 10 μg RNA from each sample was mixed to form an RNA pool (total amount 30 μg). Poly(A) mRNA was isolated and purified from the RNA pool using an Oligotex mRNA Midi kit (Qiagen, Dusseldorf, Germany). The quality and quantity of the purified RNA were determined by measuring the absorbance at 260/280-nm (A_{260}/A_{280}) using a NanoDrop ND-1000 spectrophotometer (LabTech, Hopkinton, MA USA). RNA integrity was tested by electrophoresis on 1.5% (w/v) agarose gel.

1.3 cDNA library construction and Illumina-Solexa sequencing

A random fragment sequencing library was built using a SOLiD Whole Transcriptome Analysis kit (Life Technologies, USA), in accordance with the manufacturer's standard procedure. First, the purified mRNA was fragmented in a thermomixer using the interrupt reagent and then applied as a template for first-strand cDNA synthesis. Second-strand

cDNA was synthesized using RNase H (Life Technologies), dNTP (Life Technologies), and DNA polymerase I (New England Biolabs, Ipswich, MA, USA). After purification and paired-end repair, base A was added to the 3′-end of cDNA, which was ligated to sequencing adapters. After the fragment size was selected, the fragment was amplified by polymerase chain reaction (PCR) to obtain the final sequencing library. After quality control tests using an Agilent 2100 Bioanalyzer and an ABI StepOnePlus Real-time PCR System, the library was sequenced on the Illumina HiSeq 2000 platform.

1.4 Transcriptomic data analysis

The raw reads produced by the HiSeq 2000 sequencing were cleaned by removing the adaptor sequences and empty reads, and then by filtering the low-quality reads (Phred quality<5). The clean reads were aligned to the yak genome reference sequences (version 1.0) [11] using the SOAPaligner/SOAP2 software (http://soap.genomics.org.cn/soapaligner.html). The distribution and coverage of reads on the reference genome were calculated. The gene expression levels were calculated using the RPKM method (reads per kilobase transcriptome per million mapped reads). Alternative splicing and SNP analysis were performed using the SOAPsplice software (http://soap.genomics.org.cn/soapsplice.html) and SOAPsnp software (http://soap.genomics.org.cn/soapsnp.html). The main biological functions of the mapped genes were predicted using the Blast2GO searches against the Gene Ontology (GO) database (http://www.geneontology.org/). The WEGO program (http://wego.genomics.org.cn) was used to classify the GO functions. The pathways of the mapped genes were annotated by performing local BLAST searches against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg/pathway.html). KEGG pathway enrichment analyses were performed using the GO::Term Finder program (http://smd.stanford.edu/help/GO-TermFinder/GO_TermFinder_help.shtml/) with a *P*-value ⩽0.05. A flowchart for the gene structure optimization and prediction of new transcripts is shown in Figure 1. First, the clean reads that mapped to the genome were assembled using the Cufflinks software (parameters were -u-p6-g-b-o, and for the others the default settings were used; for specific details of the settings see http://cufflinks.cbcb.umd.edu/howitworks.html). After filtering low-quality sequences (length⩽180 bp, *Q* score⩽10), the assembled sequences were compared with the annotated genes in the yak genome. If overlaps (⩾200 bp) were found on both flanks of an existing gene, the 5′ and 3′ ends of the assembled genome sequences were extended to optimize the gene structure. The assembled sequences that could not be mapped to existing genes but were located between the known genes in the genome were considered as new transcripts provided they satisfied the following requirements: the transcript must be ⩾200 bp away from an annotated gene; the transcript must be >180 bp; and the sequencing depth must be ⩾2. To analyze the function of these new transcripts further, the Coding Potential Calculator (CPC) (http://cpc.cbi.pku.edu.cn/) was used to predict their coding ability. Novel transcripts was identified using BLAST program searches against the NCBI non-redundant (nr), nucleotide (nt) and SwissProt sequence databases, and GO annotations were extracted and classified using the WEGO program.

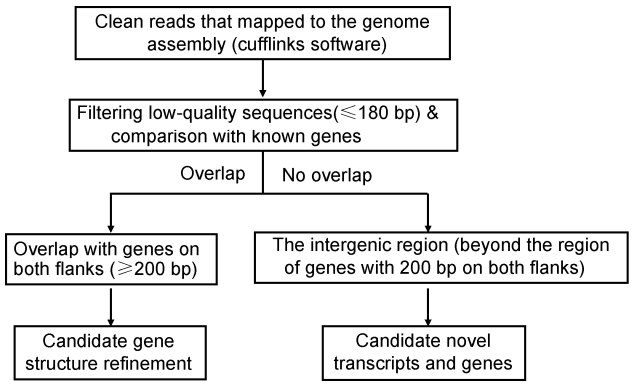

Figure 1 Gene structure refinement and candidate novel transcripts schema.

2 Results

2.1 Sequencing data analysis and annotation

The transcriptome sequencing data from yak ovary have been deposited in the NCBI Sequence Read Archive database (accession number: SRR952727). After Illumina-Solexa deep sequencing, 27749576 raw reads were obtained. The removal of low-quality reads, adaptor sequences, and empty reads resulted in 26826516 clean reads with a total of 4828772880 bp. The analyses of base composition and quality showed that the raw reads had balanced base composition (Figure 2A), and that the low-quality reads (<20) comprised a small percentage of the total reads (Figure 2B). These results indicated the quality of the raw sequencing was good. Alignment analysis showed that 33168751 (61.82%) and 16773858 (31.26%) of total clean reads (26826516) were mapped to the yak genome and to related genes, respectively. Gene coverage statistics showed that the number of genes with transcript coverage between 90% and 100% was 10295 (61%), and genes with coverage between 80% and 90% was 1623 (10%) (Figure 3). Thus, genes with high coverage accounted for the largest proportion of the mapped genes, indicating that the alignment results were good. Alternative splicing analysis showed that 3734 of the mapped genes may be involved in alternative splicing; genes involved in intron retention comprised the largest proportion, followed by genes involved in exon skipping and an alternative 3′ splice site. Based on the mapping results, a total of 81647 SNP sites were found in the

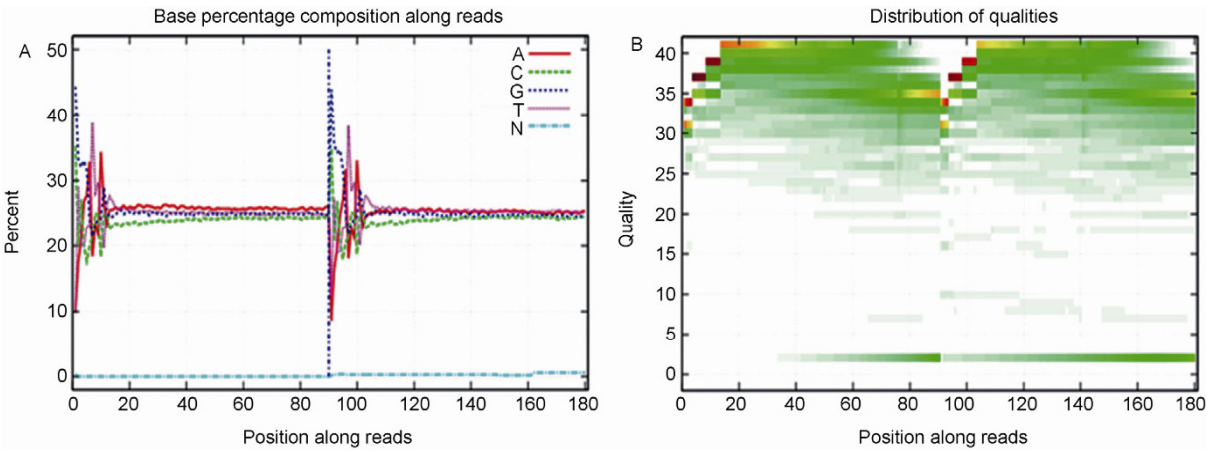

Figure 2 Base composition and quality distribution of the raw reads. A, Base composition of raw reads. On the *x*-axis, position 1–90 bp represents one read, and 91–180 bp represents the other read. The curves for the A and T bases overlap and the curves for the G and C bases overlap indicating a balanced base composition. B, Quality distribution of the bases along reads. The *x*-axis indicates the position along the reads, and the *y*-axis shows the quality value. Each dot in the image represents the quality value of the corresponding position along reads. If the percentage of the bases with low quality (<20) is low, then the sequencing quality at the corresponding position is good.

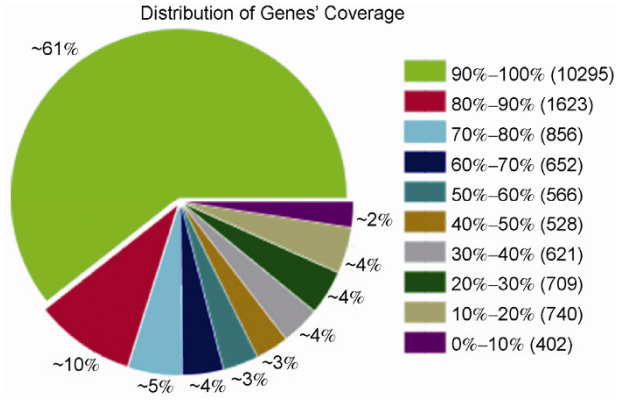

Figure 3 Coverage statistics for genes in the yak genome.

yak genome.

Analysis of the GO annotations showed that 12731 of the mapped transcripts were annotated in 59 categories under biological process, cellular component, and molecular function (Figure 4). In the biological process category, most of the annotated genes were involved in "cellular process", followed by "single-organism process" and "metabolic process". In the cellular component category, most of the genes were involved in "cell", followed by "cell part" and "organelle". In the molecular function category, most of the genes were involved in "binding", followed by "catalytic activity" and "molecular transducer activity". KEGG analysis showed that 14631 mapped transcripts were involved in 258 pathways. The top 10 enrichment pathways are shown in Table 1. Among these pathways, "focal adhesion" was the most enriched, followed by "pathways in cancer" and "extracellular matrix (ECM)-receptor interaction".

2.2 Optimization of gene structure and prediction of new transcripts

Gene structure refinement analysis showed that 7340 of the mapped genes on the yak genome could be extended based on the alignments between the transcripts and the genome sequence. Among them, 4241 genes were extended at the 5′ end and 3099 genes were extended at the 3′ end (Table S1 in Supporting Information). Prediction analysis identified 6321 new transcripts that rang in length from 180 to 14884 bp (Table S2 in Supporting Information). Exon prediction showed that the number of exon in these new transcripts ranged from 1 to 84. The CPC analysis predicted that 2267 of the new transcripts could code proteins. All the new transcripts were annotated using a local BLAST program to search against the nr, nt, and SwissProt databases. The BLAST searches aligned 4993 new transcripts to the nr database, 1453 new transcripts to the nt database, and 1200 new transcripts to the SwissProt database. The results of a statistical analysis of these new mapped transcripts are shown in Figure 5. The *E*-value distribution showed that matches with an *E*-value of 0 made up the largest portion (52.5%), followed by matches with an *E*-value of $(0–1)\times10^{-100}$ (22.9%) and 1×10^{-60}–1×10^{-45} (10.0%) (Figure 5A). The similarity distribution showed that transcripts that shared 95%–100% similarity with known sequences accounted for the largest proportion (76.3%), followed by transcripts that shared 80%–95% similarity (14.4%) and transcripts with 60%–80% similarity (4.1%) (Figure 5B). These results indicate that the BLAST results were reliable. The species distribution showed that the majority of matches were with known *Bos Taurus* sequences (41.4%), followed by *Bos grunniens mutus* (33.0%), *Ovis aries* (6.3%), *Homo sapiens* (2.8%), *Mus musculus* (1.6%), and other spe-

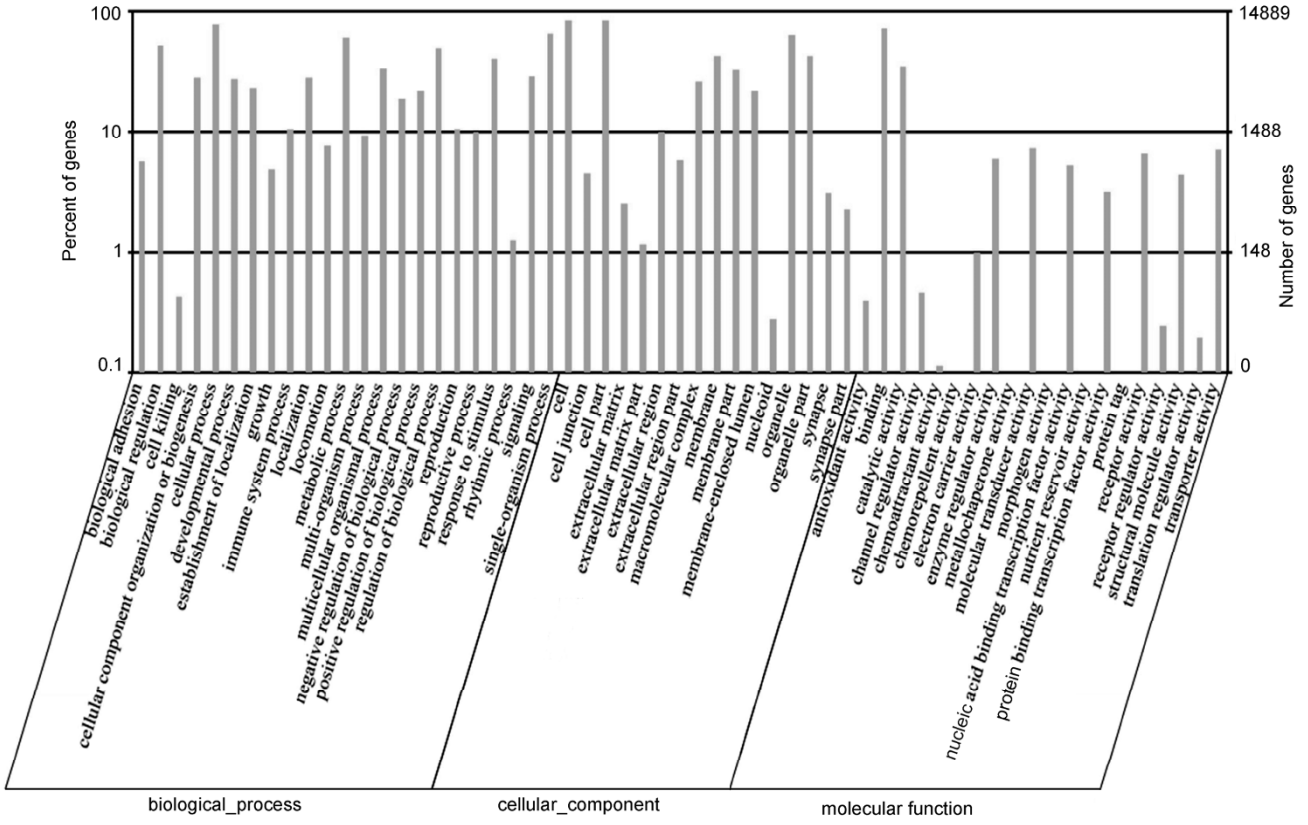

Figure 4 GO function classification of the yak ovary transcriptome.

Table 1 Top 10 pathways in KEGG enrichment analysis of the yak ovary transcriptome

	Pathway	Expressed genes with pathway annotation (14631)[a)]	All genes with pathway annotation (18965)[b)]	*P*-value
1	Focal adhesion	893 (6.1%)	985 (5.19%)	5.924522×10^{-33}
2	Pathways in cancer	559 (3.82%)	608 (3.21%)	9.409632×10^{-23}
3	ECM-receptor interaction	652 (4.46%)	720 (3.8%)	9.650817×10^{-22}
4	Amoebiasis	828 (5.66%)	938 (4.95%)	3.527666×10^{-19}
5	Regulation of actin cytoskeleton	576 (3.94%)	655 (3.45%)	6.055588×10^{-13}
6	Axon guidance	282 (1.93%)	306 (1.61%)	2.021385×10^{-12}
7	MAPK signaling pathway	303 (2.07%)	332 (1.75%)	8.067557×10^{-12}
8	Endocytosis	406 (2.77%)	455 (2.4%)	1.426242×10^{-1}
9	Insulin signaling pathway	228 (1.56%)	245 (1.29%)	1.784865×10^{-11}
10	Wnt signaling pathway	264 (1.8%)	288 (1.52%)	5.144035×10^{-1}

a) The number of expressed genes in yak ovary with pathway annotation. b) The number of genes in the whole yak genome with pathway annotation.

cies (12.8%) (Figure 5C). GO classification analysis of the new transcripts showed that the category associated with reproduction and development accounted for the largest proportion; "reproduction" was associated with the majority of the new transcripts (Table 2).

3 Discussion

RNA-Seq is a relatively efficient method for large-scale transcriptomic studies that is constantly being improved. RNA-Seq is used extensively in transcriptomic studies, including gene expression, alternative splicing, determination of non-coding RNA function, and development of SNP or SSR markers. The use of RNA-Seq in expanding gene structure information and mining of new genes has attracted considerable attention [7–10]. Yak, a typical species of herd animal distributed in the Qinghai-Tibetan Plateau, is an important production and living animal for local people. However, compared with the ordinary cattle that live in plains, yaks exhibit late sexual maturity and generally have lower fertility. The yak genome was released recently [11],

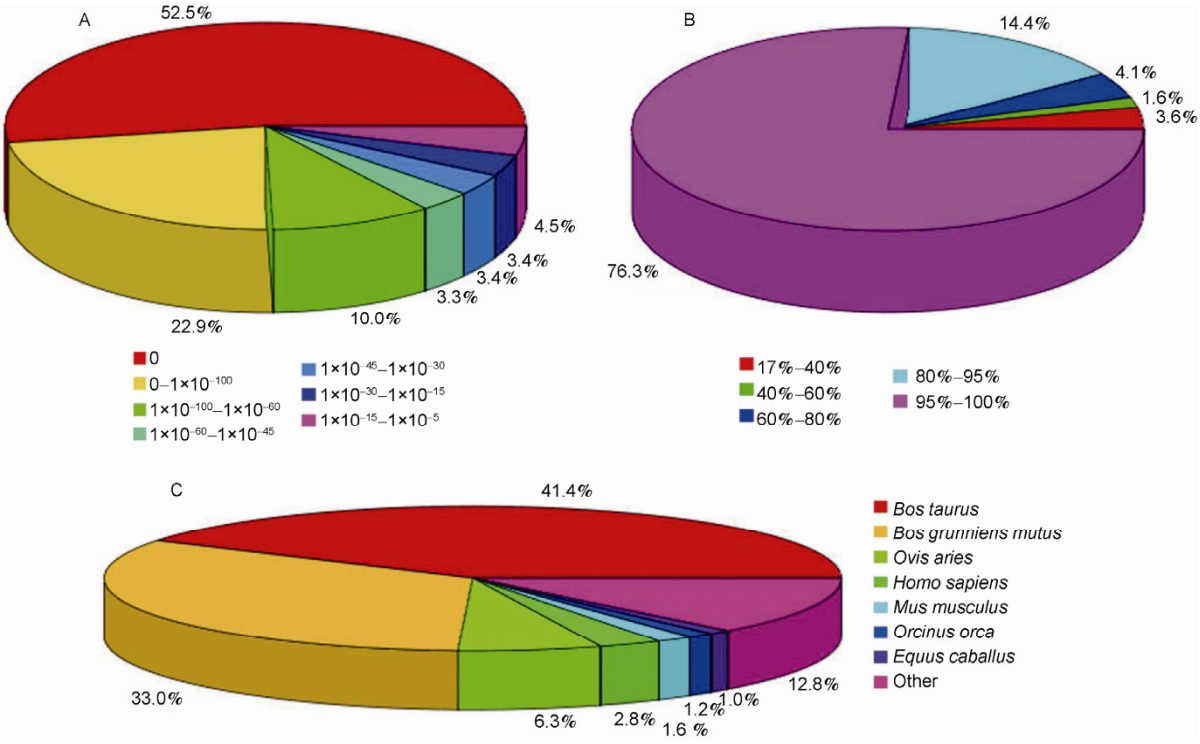

Figure 5 Alignment statistics of the new transcripts against the nr, nt, and SwissProt databases. A, *E*-value distribution. B, Similarity distribution. C, Species distribution.

Table 2 Reproduction-associated GO category assigned to the new transcripts in the yak transcriptome

	Biological process terms	Number of new transcripts	GO ID
1	reproduction	183	GO:0000003
2	reproductive process	178	GO:0022414
3	multicellular organism reproduction	93	GO:0032504
4	single organism reproductive process	91	GO:0044702
5	multi-organism reproductive process	84	GO:0044703
6	sexual reproduction	73	GO:0019953
7	developmental process involved in reproduction	63	GO:0003006
8	cellular process involved in reproduction	53	GO:0022412
9	reproductive structure development	37	GO:0048608
10	sex differentiation	35	GO:0007548
11	development of primary sexual characteristics	31	GO:0045137
12	regulation of reproductive process	29	GO:2000241
13	female pregnancy	16	GO:0007565
14	ovulation cycle process	15	GO:0022602
15	female sex differentiation	15	GO:0046660
16	ovarian follicle development	6	GO:0001541
17	oocyte differentiation	6	GO:0009994
18	oocyte development	6	GO:0048599
19	sex determination	4	GO:0007530
20	embryo implantation	4	GO:0007566
21	oocyte maturation	2	GO:0001556
22	multicellular organismal reproductive behavior	2	GO:0033057

thereby providing an important foundation for the understanding of yak characteristics at the molecular level. To date, no study related to the yak transcriptome has been reported. In the present study, RNA-Seq HT technology was employed to analyze the transcriptome of yak ovary. The normal transcriptome map of the yak ovary is described, which will contribute to the further improvement of yak gene structural information and mining of potential new genes.

3.1 Genetic architecture of the normal yak ovary transcriptome

In this study, a yak ovary RNA pool was constructed by mixing three mature healthy female yak ovary samples, to obtain an ovary sequencing library, which after sequencing, yielded 27749576 raw reads. After removing low-quality sequences, 26826516 clean reads (4 Gb data) were obtained. The base composition and quality analyses showed that the ratio of Q20 (i.e., the quality of bases ⩾20) was 94.7% and the GC content was 45.5%, which indicated successful library construction and good sequencing quality. Alignment analysis showed that 16992 yak genes were mapped, of which 3734 mapped genes were involved in alternative splicing. The 16992 mapped genes were identified as transcripts in the normal mature yak ovary. No related study and data for the ovary transcriptome of other bovine species have been reported. However, gene expression of bovine oocytes and granule cells has been investigated by DNA microarray. A total of 13162 and 13602 genes were reported to be expressed in different development stages of oocytes and granule cells, respectively [12], and a total of 8489 genes were found to be expressed at different stages of bovine oocytes development [13]. In the present study, the data generated were more comprehensive than in the previous studies, suggesting that the number of genes involved in the normal physiological functions and development of the ovary is actually significantly more than was revealed by the DNA microarray results.

The functions of the expressed genes were annotated by GO and KEGG analyses. A total of 23 GO categories under biological processes were assigned to the transcripts, of which the largest proportion was "cellular process", followed by "single-organism process" and "metabolic process". In addition, 7068 genes were associated with development and reproduction, namely, "developmental and reproductive process", "reproduction", and "reproductive process". These results are consistent with the biological characteristics and function of the ovary. A total of 16 GO categories under cellular component were assigned to the transcripts, of which the largest proportion was "cell-related part". In addition, a significant proportion of the transcripts were associated with membrane component, indicating that membrane components are important in the physiological activity of the yak ovary. A total of 20 GO categories under molecular function were assigned to the transcripts, of which the largest proportion was "binding". Previous microarray studies showed that the RNA-binding molecular function category accounted for a large proportion in the expressed genes in bovine oocytes [12,13]; thereby confirming the results of the present study that "binding" plays an important role in the normal physiological activities of the yak ovary.

KEGG analysis predicted that the expressed genes were involved in 258 pathways, of which "focal adhesion" was the most enriched. Focal adhesion, a connection function mediated by cells and ECM, results in a dynamic cell anchor type of connection, in which the integrins anchor to the ECM [14]. In addition to the "focal adhesion" pathway, the "ECM-receptor interaction" pathway was also among the top 10 pathways. The ECM is a complex matrix of biological macromolecules, such as glycoproteins, protein polysaccharides, and amino sugars. The ECM has an important function in various aspects of cell physiological activities, including cell adhesion, movement, proliferation, and differentiation, by interactions with its surface receptor (ECM receptor). For example, in cell adhesion, the ECM can transfer signals to the cells via the surface receptors, and through various signaling transduction pathways, the ECM can send signals to the cytoplasm and nucleus to influence gene expression or cellular activities [15]. In the present study, transcripts associated with the "ECM-receptor interaction" pathway were highly expressed, which suggested that they may be complementary to "focal adhesion," and both pathways have an important functions in promoting cell adhesion and connection. The enrichment of these two pathways in the transcriptome indicated that cell connections occur extensively in the yak ovary. Previous studies have shown that the follicles of ovary do not have a micro-environment vascular system. In addition, a wide gap connection exists between oocytes, cumulus cells, and other follicular cells, thereby forming a complete functional joint venture. An oocyte mainly communicates with its surrounding cells (granulose and theca cells) through cell adhesion and connection. Numerous small molecular substances that contain information, nutrients, and metabolites that regulate oocyte growth and development are transported through this connection [16,17]. The results from the present study confirmed that adhesion and connection mechanisms are important in the physiological activity of ovary at the molecular level. The "regulation of actin cytoskeleton" pathway was also enriched in the transcriptome. The actin cytoskeleton is a dynamic skeletal cell structure that contains actin and associated proteins, which has important functions in various physiological processes, ranging from movement to membrane transport [18]. Recent studies have shown that actin cytoskeleton also has an important function in the early development of oocyte organization and maturation [19], by assisting in the cytokinesis or the formation of cytoplasmic channels, as well as in the transport of oocyte-specific RNA and proteins. The high expression of transcripts involved in the "regulation of actin cytoskeleton" pathway suggested the existence of dynamic changes in the cytoskeleton structure of yak ovary. Gamete maturation is a complex process, which involves numerous spatial and morphological regulations. We speculate that this organization may be a dynamic spatial adjustment, which is related to oocyte maturation and development of other reproductive functions in the ovary [20].

Among the top 10 pathways, three signaling pathways,

mitogen-activated protein kinase (MAPK), insulin, and Wnt, were also enriched. The role of the MAPK and Wnt signaling pathways in oocyte maturation has been well recognized. MAPK is a serine/threonine protein kinase that is available in various signaling pathways, which acts as a common component of signal transduction in the regulation of cell growth and plays an important role in cell growth and differentiation as well as in the cell cycle. The MAPK signaling pathway is also important in eukaryotic signaling networks where it plays a role in passing the upstream signal to the downstream elements [21]. The MAPK pathway is activated when phosphorylation occurs during oocyte maturation; therefore, the MAPK pathway is important in oocyte maturation and mature metaphase II arrest [19,22]. Wnt is a secreted glycoprotein, which participates in autocrine or secretions. The Wnt protein is important in the regulation of cell proliferation, differentiation, and migration during an organism's growth and development, and it can determine cell polarity, fate, and proliferation of progenitor cells [23]. Recent studies have shown that the Wnt signaling pathway is necessary to regulate the normal development of the mammalian reproductive system. This pathway is involved mainly in the formation of Mueller's pipe, control of follicular development, ovulation, and luteinization, and as well as in the establishment of normal pregnancy [24–26]. The enrichment of the MAPK and Wnt pathways in the yak ovary transcriptome is similar to previous studies on oocytes of different species [19,22,24,26], thereby indicating that these two signaling pathways are important in maintaining physiological activity of the ovary.

The finding that the insulin signaling pathway was also enriched in the yak ovary transcriptome is novel and has not been reported previously. Insulin is a multifunctional protein peptide that can induce a series of signal cascade reactions by combining with its insulin-specific receptor. Two major signal transduction pathways are involved in the insulin signaling pathway; one is the phosphatidyl inositol 3-kinase/protein kinase B pathway, which mainly regulates cell metabolism, survival, and apoptosis, and the other is the MAPK/extracellular signal-regulated kinase pathway, which is important in the regulation of embryonic development as well as in cell differentiation, proliferation, and apoptosis [27]. Recent studies have confirmed that insulin is an important factor in promoting proliferation of ovarian granulose cell and other related physiological functions that regulate the proliferation of granulose cells, as well as the metabolism and synthesis of steroid hormones [28,29]. Insulin-related disorders such as polycystic ovary syndrome, occur in humans, which results in the inhibition of granular cell proliferation and dysfunctional or inhibition of follicular growth and ovulation [30]. In the present study, all the genes that are involved in insulin signaling pathway were significantly expressed in the transcriptome, and two principal signal transduction pathways were activated (Figure 6). This result suggests that the insulin signaling pathway is important in the physiological activity of yak ovary, where the insulin signaling pathway may regulate not only energy metabolism, but also cell differentiation and apoptosis. This pathway can integrate the energy metabolic and reproductive activities of the ovary.

Three of the top 10 enriched pathways, namely, "pathways in cancer", "amoebiasis", and "axon guidance", have no obvious associations with the reproductive function of the ovary. Although these pathways were derived from their corresponding physiological processes, we speculate that some of the genes involved in these pathways may be related to the functions and activities of the ovary. For example, although "pathway in cancer" is related to cancer, an important feature of cancer is proliferation, which is related to the proliferation of germ cells. Thus, some genes involved in the "pathway in cancer" may also be involved in the proliferation and development of ovary-related cells. However, further studies are needed to determine the exact functions of these pathways in the yak ovary.

3.2 Optimization of yak gene structure and discovery of new transcripts

With the completion of the human genome sequencing, the genome sequences of many other species have also been completed. However, the majority of the current gene annotations have been predicted based on the existing model sequences using bioinformatics tools. Small molecular peptides, lowly expressed genes, or species-specific genes cannot be predicted using these tools. Thus, some of the gene annotations for the reference genomes may be incomplete or missing. For example, although the human genome map has been publicly available for many years, the discovery of new genes and the optimization of the original gene annotation are still continually being developed [31–39]. In earlier studies, researchers built EST libraries and used electronic cloning or SAGE technology based on the available EST sequences to mine for new genes [40–46]. These methods have low efficiency and are not sufficiently comprehensive. Recently, RNA-Seq has been used extensively in various transcriptomic studies. For example, Chen et al. [7] used RNA-Seq to analyze the transcriptome of many tissues from two hybrid pigs (White Duroc×Erhualian). They found that 2012 and 2083 genes could be extended at the 5′ ends, 3471 and 3648 could be extended at the 3′ ends, and 1655 and 1637 could be extended at both the 5′ and 3′ends, which helped them obtain more accurate structures for many genes. Lu et al. [8] analyzed two subspecies of cultivated rice using RNA-Seq and identified 15708 new transcriptionally active areas. They found 6228 genes that could be extended at both the 5′ and 3′ ends. Jäger et al. [9] compared the gene expression profiles of normal sheep tissue with those of delayed bone healing tissue using RNA-Seq and identified 12431 new transcripts. Huang et al. [10] used RNA-Seq to analyze the transcriptome of bovine embryos at different

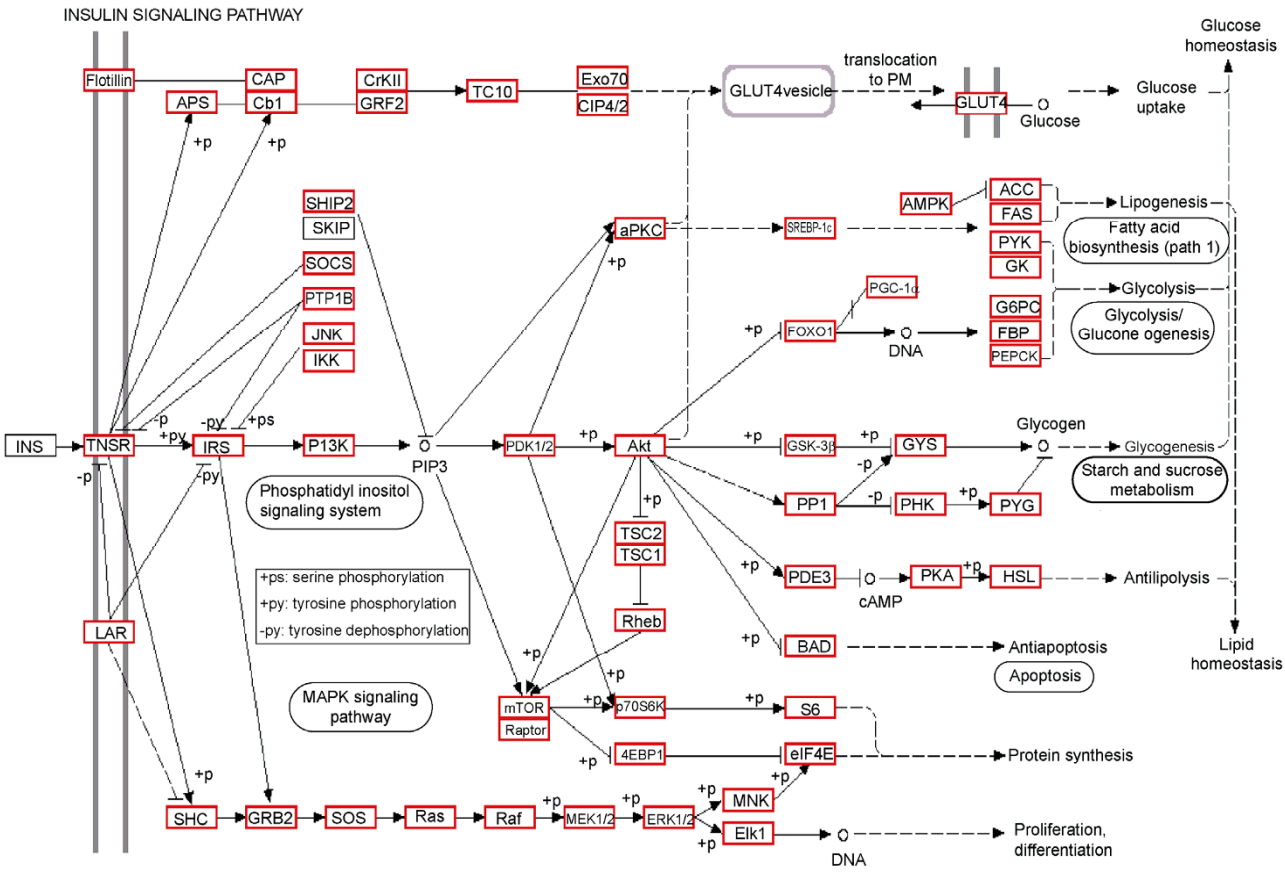

Figure 6 Genes in the insulin signaling pathway significantly expressed in the yak transcriptome. Red boxes indicate the genes that were expressed in the transcriptome.

developmental stages. They identified 1785 new transcripts after comparing their results with the bovine genome.

The yak genome was released recently [11], bringing yak studies into the post-genomic era. The genome information should now be improved and potential new genes should be mined. In the present study, RNA-Seq was employed to analyze the yak ovary transcriptome. After aligning the reads to the yak genome, 7340 of the originally predicted genes cloud be extended at their 5′ or 3′ ends, which indicated that the original predictions of 5′ or 3′ UTRs could contain errors. These extensions could be used to optimize the structure of the genes in the genome. For the assembled reads that were not mapped to existing genes but that were located between them in the genome, further analysis was performed. The results showed that 6321 new transcripts, with lengths ranging from 180 to 14884 bp, were obtained. Exon prediction, which was performed by locating genes in the genome, showed that the number of enxon in these new transcripts ranged from 1 to 84. The CPC analysis revealed that 2267 of these new transcripts can code proteins. The results provide valuable data for mining new yak genes. The new transcripts were annotated using local BLAST searches against the nr, nt, and SwissProt databases. The results show that 4993, 1453, and 1200 new transcripts were mapped to the nr, nt, and SwissProt databases, respectively, which further confirmed the presence of new genes in the yak genome. Comparative statistical analysis of the mapped new transcripts showed that majority of matches were to *B. taurus* (41.4%) sequences, followed by *B. grunniens mutus* (33.0%), *O. aries* (6.3%), *H. sapiens* (2.8%), *M. musculus* (1.6%), and other species. The transcripts that shared high similarity with bovine genes accounted for the largest proportion of the potential new yak genes. Transcripts that shared high similarity with *B. grunniens mutus* genes also account for a large proportion of potential new yak genes. Because the local yak (*B. grunniens*) was used as the source of DNA for the genome sequencing [11], we speculate that many gaps exist in the current genome annotation and that genes that are similar to *B. grunniens mutus* genes have not been annotated. A considerable proportion of the potential new genes were also similar to other species; therefore, further studies are required to improve the yak genome annotations. GO analysis of the new transcripts showed that the category associated with reproduction and development accounted for the largest proportion, with the majority of them being involved in the “reproduction” category. These results provide robust data for further mining and annotation of novel genes that are related to yak reproduction.

In summary, RNA-Seq HT technology was employed to analyze the transcriptome of yak ovary. The results provide a gene expression pattern for normal yak ovary as a basis for future studies on yak breeding performance. RNA-Seq provided valuable data that were used to improve the yak genome structure information and to mine for new genes.

This work was supported by the National Science & Technology Pillar Program of China (2012BAD13B06) and the Special Fund for Agroscientific Research in the Public Interest (201203009).

1 Suzuki Y, Sugano S. Transcriptome analyses of human genes and applications for proteome analyses. Curr Protein Pept Sci, 2006, 7: 147–163
2 Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y, Carninci P. The complexity of the mammalian transcriptome. J Physiol, 2006, 575: 321–332
3 Sangwan RS, Tripathi S, Singh J, Narnoliya LK, Sangwan NS. De novo sequencing and assembly of centella asiatica leaf transcriptome for mapping of structural, functional and regulatory genes with special reference to secondary metabolism. Gene, 2013, 525: 58–76
4 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
5 Tariq MA, Kim HJ, Jejelowo O, Pourmand N. Whole-transcriptome RNASeq analysis from minute amount of total RNA. Nucleic Acids Res, 2011, 39: e120
6 Richard H, Schulz MH, Sultan M, Nurnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, Haas SA, Yaspo ML. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Res, 2010, 38: e112
7 Chen C, Ai H, Ren J, Li W, Li P, Qiao R, Ouyang J, Yang M, Ma J, Huang L. A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. BMC Genomics, 2011, 12: 448
8 Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, Han B. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-Seq. Genome Res, 2010, 20: 1238–1249
9 Jager M, Ott CE, Grunhagen J, Hecht J, Schell H, Mundlos S, Duda GN, Robinson PN, Lienau J. Composite transcriptome assembly of RNA-Seq data in a sheep model for delayed bone healing. BMC Genomics, 2011, 12: 158
10 Huang W, Khatib H. Comparison of transcriptomic landscapes of bovine embryos using RNA-Seq. BMC Genomics, 2010, 11: 711
11 Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, Auvil L, Capitanu B, Ma J, Lewin HA, Qian X, Lang Y, Zhou R, Wang L, Wang K, Xia J, Liao S, Pan S, Lu X, Hou H, Wang Y, Zang X, Yin Y, Ma H, Zhang J, Wang Z, Zhang Y, Zhang D, Yonezawa T, Hasegawa M, Zhong Y, Liu W, Huang Z, Zhang S, Long R, Yang H, Lenstra JA, Cooper DN, Wu Y, Shi P, Liu J. The yak genome and adaptation to life at high altitude. Nat Genet, 2012, 44: 946–949
12 Regassa A, Rings F, Hoelker M, Cinar U, Tholen E, Looft C, Schellander K, Tesfaye D. Transcriptome dynamics and molecular cross-talk between bovine oocyte and its companion cumulus cells. BMC Genomics, 2011, 12: 57
13 Mamo S, Carter F, Lonergan P, Leal CL, Al Naib A, McGettigan P, Mehta JP, Evans AC, Fair T. Sequential analysis of global gene expression profiles in immature and *in vitro* matured bovine oocytes: potential molecular markers of oocyte maturation. BMC Genomics, 2011, 12: 151
14 Chen CS, Alonso JL, Ostuni E, Whitesides GM, Ingber DE. Cell shape provides global control of focal adhesion assembly. Biochem Biophys Res Commun, 2003, 307: 355–361
15 Alldinger S, Groters S, Miao Q, Fonfara S, Kremmer E, Baumgartner W. Roles of an extracellular matrix (ECM) receptor and ecm processing enzymes in demyelinating canine distemper encephalitis. Dtsch Tierarztl Wochenschr, 2006, 113: 151-152, 154–156
16 Eppig JJ. Intercommunication between mammalian oocytes and companion somatic cells. Bioessays, 1991, 13: 569–574
17 Gilchrist RB, Ritter LJ, Armstrong DT. Oocyte-somatic cell interactions during follicle development in mammals. Anim Reprod Sci, 2004, 82–83: 431–446
18 Saarikangas J, Zhao H, Lappalainen P. Regulation of the actin cytoskeleton-plasma membrane interplay by phosphoinositides. Physiol Rev, 2010, 90: 259–289
19 Sun QY, Schatten H. Regulation of dynamic events by microfilaments during oocyte maturation and fertilization. Reproduction, 2006, 131: 193–205
20 Brunet S, Maro B. Cytoskeleton and cell cycle control during meiotic maturation of the mouse oocyte: integrating time and space. Reproduction, 2005, 130: 801–811
21 Seger R, Krebs EG. The mapk signaling cascade. FASEB J, 1995, 9: 726–735
22 Harrouk W, Clarke HJ. Mitogen-activated protein (MAP) kinase during the acquisition of meiotic competence by growing oocytes of the mouse. Mol Reprod Dev, 1995, 41: 29–36
23 Huang H, He X. Wnt/beta-catenin signaling: new (and old) players and new insights. Curr Opin Cell Biol, 2008, 20: 119–125
24 Harwood BN, Cross SK, Radford EE, Haac BE, De Vries WN. Members of the Wnt signaling pathways are widely expressed in mouse ovaries, oocytes, and cleavage stage embryos. Dev Dyn, 2008, 237: 1099–1111
25 Boyer A, Goff AK, Boerboom D. Wnt signaling in ovarian follicle biology and tumorigenesis. Trends Endocrinol Metab, 2010, 21: 25–32
26 Wang HX, Tekpetey FR, Kidder GM. Identification of Wnt/beta-catenin signaling pathway components in human cumulus cells. Mol Hum Reprod, 2009, 15: 11–17
27 Woods YL, Petrie JR, Sutherland C. Dissecting insulin signaling pathways: individualised therapeutic targets for diagnosis and treatment of insulin resistant states. Endocr Metab Immune Disord Drug Targets, 2009, 9: 187–198
28 Seto-Young D, Zajac J, Liu HC, Rosenwaks Z, Poretsky L. The role of mitogen-activated protein kinase in insulin and insulin-like growth factor I (IGF-I) signaling cascades for progesterone and IGF-binding protein-1 production in human granulosa cells. J Clin Endocrinol Metab, 2003, 88: 3385–3391
29 Richardson MC, Cameron IT, Simonis CD, Das MC, Hodge TE, Zhang J, Byrne CD. Insulin and human chorionic gonadotropin cause a shift in the balance of sterol regulatory element-binding protein (SREBP) isoforms toward the SREBP-1c isoform in cultures of human granulosa cells. J Clin Endocrinol Metab, 2005, 90: 3738–3746
30 Fornes R, Ormazabal P, Rosas C, Gabler F, Vantman D, Romero C, Vega M. Changes in the expression of insulin signaling pathway molecules in endometria from polycystic ovary syndrome women with or without hyperinsulinemia. Mol Med, 2010, 16: 129–136
31 Duchateau PN, Pullinger CR, Cho MH, Eng C, Kane JP. Apolipoprotein l gene family: tissue-specific expression, splicing, promoter regions; discovery of a new gene. J Lipid Res, 2001, 42: 620–630
32 Jia HP, Schutte BC, Schudy A, Linzmeier R, Guthmiller JM, Johnson GK, Tack BF, Mitros JP, Rosenthal A, Ganz T, McCray PB, Jr. Discovery of new human beta-defensins using a genomics-based approach. Gene, 2001, 263: 211–218
33 Whitney G, Wang S, Chang H, Cheng KY, Lu P, Zhou XD, Yang WP, McKinnon M, Longphre M. A new siglec family member, siglec-10, is expressed in cells of the immune system and has signaling properties similar to CD33. Eur J Biochem, 2001, 268: 6083–6096
34 Bourdon V, Naef F, Rao PH, Reuter V, Mok SC, Bosl GJ, Koul S, Murty VV, Kucherlapati RS, Chaganti RS. Genomic and expression analysis of the 12p11-p12 amplicon using est arrays identifies two novel amplified and overexpressed genes. Cancer Res, 2002, 62:

6218–6223
35 Kao CY, Chen Y, Zhao YH, Wu R. Orfeome-based search of airway epithelial cell-specific novel human β-defensin genes. Am J Respir Cell Mol Biol, 2003, 29: 71–80
36 Seibold S, Rudroff C, Weber M, Galle J, Wanner C, Marx M. Identification of a new tumor suppressor gene located at chromosome 8p21.3-22. FASEB J, 2003, 17: 1180–1182
37 Li D, Lu GX, Fu JJ, Mo YQ, Xing XW, Liu G. Molecular cloning and expression analysis of a novel human testis-specific gene (in Chinese). Yi Chuan Xue Bao, 2004, 31: 545–551
38 Guo LL, Ci HL, Shan HS, Zou X, Zhai YG, Li YP. Molecular cloning and expression analysis of a novel human gene znf18 (in Chinese). Yi Chuan, 2005, 27: 523–530
39 Harding MA, Theodorescu D. Rhogdi2: a new metastasis suppressor gene: discovery and clinical translation. Urol Oncol, 2007, 25: 401–406
40 Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y. Normalization and subtraction of cap-trapper-selected cdnas to prepare full-length cdna libraries for rapid discovery of new genes. Genome Res, 2000, 10: 1617–1630
41 Huminiecki L, Bicknell R. *In silico* cloning of novel endothelial-specific genes. Genome Res, 2000, 10: 1796–1806
42 Schultz J, Doerks T, Ponting CP, Copley RR, Bork P. More than 1000 putative new human signalling proteins revealed by EST data mining. Nat Genet, 2000, 25: 201–204
43 Nagase T, Nakayama M, Nakajima D, Kikuno R, Ohara O. Prediction of the coding sequences of unidentified human genes. XX. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*. DNA Res, 2001, 8: 85–95
44 Wittenberger T, Schaller HC, Hellebrand S. An expressed sequence tag (EST) data mining strategy succeeding in the discovery of new G-protein coupled receptors. J Mol Biol, 2001, 307: 799–813
45 Boheler KR, Stern MD. The new role of sage in gene discovery. Trends Biotechnol, 2003, 21: 55–57; discussion 57–58
46 Dalla E, Mignone F, Verardo R, Marchionni L, Marzinotto S, Lazarevic D, Reid JF, Marzio R, Klaric E, Licastro D, Marcuzzi G, Gambetta R, Pierotti MA, Pesole G, Schneider C. Discovery of 342 putative new genes from the analysis of 5′-end-sequenced full-length-enriched cDNA human transcripts. Genomics, 2005, 85: 739–751

Supporting Information

Table S1 Extended genes in the yak genome

Table S2 New transcripts in the yak transcriptome

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.