

# Identification and Functional Analysis of Transcriptome Profiles, Long Non-coding RNAs, Single-Nucleotide Polymorphisms, and Alternative Splicing from the Oocyte to the Preimplantation Stage of Sheep by Single-Cell RNA Sequencing

[Zi Jing Zhang](#) , Qiao ting Shi , Xiao ting Zhu , Lei Jin , Li min Lang , [Shi jie Lyy](#) , Xiao ling Xin , [Yong zhen Huang](#) , Xiang Yu , Zhi ming Li , [Su Juan Chen](#) , [Zhao xue Xu](#) , [Wei Zhang](#) \* , [Er yao Wang](#) \*

Posted Date: 23 April 2023

doi: 10.20944/preprints202304.0773.v1

Keywords: Key words: transcriptome profiles; long non-coding RNAs; single-nucleotide polymorphisms; alternative splicing; sheep; preimplantation; Single-cell RNA sequencing



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# Identification and Functional Analysis of Transcriptome Profiles, Long Non-Coding RNAs, Single-Nucleotide Polymorphisms, and Alternative Splicing from the Oocyte to the Preimplantation Stage of Sheep by Single-Cell RNA Sequencing

Zijing Zhang <sup>1</sup>, Qiaoting Shi <sup>1</sup>, Xiaoting Zhu <sup>1</sup>, Lei Jin <sup>1</sup>, Limin Lang <sup>1</sup>, Shijie Lyv <sup>1</sup>, Xiaoling Xin <sup>1</sup>, Yongzhen Huang <sup>2</sup>, Xiang Yu <sup>4</sup>, Zhiming Li <sup>3</sup>, su Juan Chen <sup>5</sup>, Zhaoxue Xu <sup>1</sup>, Wei Zhang <sup>6,\*</sup> and Eryao Wang <sup>1,\*</sup>

<sup>1</sup> Institute of Animal Husbandry and Veterinary Science, Henan Academy of Agricultural Sciences, Zhengzhou, Henan, 450002, People's Republic of China; Zijing Zhang: vincezhang163@163.com; Qiaoting: sqts@126.com; Xiaoting Zhu: xtzh@2019@outlook.com; Lei Jin: jinlei940304@126.com; Limin Lang: langlimin666@outlook.com; Shijie Lyv: sjly@outlook.com; Xiaoling Xin: xinxl726@163.com; Zhaoxue Xu: xuzhaoxue11@163.com;

<sup>2</sup> College of Animal Science and Technology, Northwest A&F University, Yangling Shaanxi, 712100, People's Republic of China; Yongzhen Huang: hyzsci@nwfau.edu.cn

<sup>3</sup> Henan Provincial Animal Husbandry General Station, Zhengzhou, Henan, 450008, People's Republic of China; Zhiming Li: hna1920@163.com

<sup>4</sup> Henan Animal Health Supervision Institute Zhengzhou, Henan, 450003, People's Republic of China; Xiang Yu: 15838282287@163.com

<sup>5</sup> Synthetic Biology Engineering Lab of Henan Province, School of Sciences and Technology, Xinxiang Medical University, Henan, 453003, People's Republic of China; Chen su Juan: chensujuan101@163.com

<sup>6</sup> College of Animal Science and Veterinary Medicine, Henan Institute of Science and Technology, Xinxiang, 453003, People's Republic of China

\* Correspondence: yuntaishan2003@aliyun.com (W.Z.); wangeryao666@qq.com (E.W.); Tel.: +86037165727015; Fax: +86037165727015; Address: NO.116, Hua yuan Road, Zhengzhou, Henan, People's Republic of China

**Abstract:** Numerous dynamic and complicated processes characterize development from the oocyte to the embryo. However, given the importance of functional transcriptome profiles, long non-coding RNAs, single-nucleotide polymorphisms, and alternative splicing during embryonic development, the effect that these features have on the blastomeres of 2-, 4-, 8-, 16-cell, and morula stages of development have not been studied. Here, we conducted a scRNA-seq survey of cells from sheep from the oocyte to the blastocyst developmental stages. We then carried out experiments to identify and functionally analyze the transcriptome profiles, long non-coding RNAs, single-nucleotide polymorphisms (SNPs), and alternative splicing (AS). We founded that between the oocyte and zygote groups significantly down-regulated genes and the second-largest change in gene expression occurred between the 8- and 16-cell stages. We used various methods to construct a profile to characterize cellular and molecular features and systematically analyze the related GO and KEGG profile of cells of all stages from the oocyte to the blastocyst. This large-scale, single-cell atlas provides key cellular information and will likely assist clinical studies in improving preimplantation genetic diagnosis.

**Keywords:** transcriptome profiles; long non-coding RNAs; single-nucleotide polymorphisms; alternative splicing; sheep; preimplantation; single-cell RNA sequencing

## 1. Introduction

From the development of the oocyte to the embryo, many dynamic and complicated processes take place as the zygote undergoes several rapid rounds of division and produces a mass of cells within the zona pellucida [1,2]. During these dynamic stages, differential gene expression in individual

cells is a key determinant of cellular differentiation, function, and physiology [3,4]. In recent years, several studies have documented the key developmental processes underlying the formation of blastomeres of 4- and 8-cell embryos of mice [5], cattle [6], and goats [7]. However, the exact mechanism and developmental patterns underlying how the blastomeres of the 2-, 4-, 8-, 16-cell, morula, and blastocyst stages undergo asymmetric division are still unclear. Understanding the molecular mechanism underlying cleavage-stage development is critically important for improving preimplantation genetic diagnosis.

Single-cell RNA sequencing (scRNA-seq) provides an alternative method for studying the cellular heterogeneity of human preimplantation embryos and embryonic stem cells [8,9], mouse oocytes [10], porcine oocyte maturation [11], and Haimen white goat's oocyte maturation by generating a readout of the abundance of a transcript within a cell [12]. Indeed, scRNA-seq applied to mammalian gametes has generated new insights into the composition of transcripts and disease-related functional abnormalities. However, scRNA-seq studies have not thoroughly characterized how the blastomeres of the 2-, 4-, 8-, 16-cell, and especially the morula and blastocyst stages, undergo asymmetric division. Here, we aimed to transcriptionally profile nucleated cells present during the blastomeres of 2-, 4-, 8-, 16-cell and morula stage undergoing asymmetric division to provide a broad profile of composition of transcripts in the cell and their transcriptomes.

Long non-coding RNAs (lncRNAs) are a diverse group of RNAs that are longer than an arbitrary limit of 200 nt and do not encode proteins [13]. Nevertheless, lnc RNAs can be located in exonic, intronic, and intergenic regions and can regulate gene expression by interacting with other biological macromolecules, such as RNA, DNA, proteins, and other factors, to promote normal cell function [14,15]. Compared with the characteristics of protein-coding genes, lncRNAs tend to be less conserved across species and often show lower expression levels and high tissue specificity [16]. During the development of embryos, some lncRNAs, such as Xist, Tsix, and H19, have significant regulatory functions and can potentially determine the cell's fate and direction of differentiation during embryogenesis to form different organs or special tissues that contain various cells expressing specific genes [17–19].

An increasing number of studies have reported several thousands of annotated lncRNA loci in mammalian oocytes and early embryos [1,20]. For example, a total of 2733 novel lncRNAs were found to be expressed in specific developmental stages among 8701 lncRNAs using single-cell sequencing analysis of 124 individual cells from human embryonic stem cell (ESCs) and human preimplantation embryos at different passages. In addition, 5204 novel lncRNAs were obtained from in vivo and somatic cell nuclear transfer mouse preimplantation embryo, suggesting that several lncRNAs and their association with known protein-coding genes might be involved in mouse embryonic development and cell reprogramming [8,21]. Another study reported that approximately a quarter of 1600 lncRNA loci expressed during the murine oocyte-to-embryo transition employed promoters from a long terminal repeat retrotransposon class, which exhibited either maternal or zygotic expression and showed signatures of massive expansion in the evolution of the rodent genome [22]. In bovine early embryos, some lncRNAs play a role in the translational control of target mRNA and are thus important for managing maternal reserves, especially before embryonic genome activation, as these reserves contain the embryonic program [23]. Despite the fact that various functional lncRNAs are important during embryonic development, the effect of lncRNAs on the blastomeres of the 2-, 4-, 8-, 16-cell, and morula stages of development has not been studied; nevertheless, this subject is in need of more attention and discussion by scientists. lncRNAs play an important role in biological processes, including epigenetic regulation, dose compensation, cell cycle, cell differentiation, proliferation, apoptosis through gene imprinting, chromatin remodeling, transcriptional activation, transcriptional interference, and nuclear splicing regulation.

Here, we conducted an scRNA-seq survey of 24 sheep cells during their development from the oocyte to the blastocyst stages. We then conducted experiments involved in the identification and functional analysis of transcriptome profiles, lncRNAs, single-nucleotide polymorphisms (SNPs) and alternative splicing (AS). Using these different methods, we constructed a profile to characterize cellular and molecular features and systematically analyze the related GO and KEGG profile of cells

during sheep development. This large-scale single-cell atlas provides key cellular information and will likely aid clinical studies in the development of more efficient preimplantation genetic diagnosis.

## 2. Materials and Methods

### 2.1. Animals and sample collection

All work with animals was completed in accordance and with the approval of the Henan Academy of Agricultural Sciences Institutional Animal Care and Use Committee. Mature sheep were obtained from Hui yuan Sheep Industry Co., Ltd (pu yang, he nan province). We used 15-month-old female sheep (40 kg) for our study. The animals were provided with grass and drinking water and had access to an animal exercise pen. All animals were healthy, showed normal appetite, and had smooth wool. Artificial insemination using semen from one of three fertile bulls was conducted at 12- and 24-hours post-standing heat (Day 0). Donor animals were sacrificed at 30 hours, and in vivo developed oocytes and embryos at the 2- to 16-cell stages were collected by oviductal flushing 2–4 days after estrus. Early morulae, compact morulae, and blastocysts were collected by routine non-surgical uterine flushing on days 5, 6, and 7. All oocytes and embryos were examined and staged under light microscopy. Only morphologically intact embryos meeting the standards of Grade 1 by the International Embryo Transfer Society were included in this study. Embryos were washed twice in D-PBS before being frozen and stored individually in RNAlater (Ambion, Grand Island, NY, USA) in liquid nitrogen.

### 2.2. RNA isolation, library preparation and sequencing

Firstly, total RNA was isolated from the sheep sample (three biological replicates per sample combined) using single cell RNA kit (Single Cell RNA Purification Kit NGB) was extracted (Norgen Biotek, CA) following the manufacturer's procedure. The single cell RNA quantity and purity were analysis of Bioanalyzer 2100 and RNA 6000 Nano LabChip Kit (Agilent, CA, USA) with RIN number >7.0. Approximately 10 ug of total RNA representing a specific adipose type was subjected to isolate Poly (A) mRNA with poly-T oligoattached magnetic beads (Invitrogen). Following purification, all amplifications were carried out in parallel with positive and no-template controls for quality assurance which using the SMARTer Universal low Input RNA kit (Clontech) for cDNA library. Briefly, RNA is converted to cDNA, and then adapters for Illumina sequencing (with specific barcodes) are added through PCR using only a limited number of cycles. The PCR products are purified (Protocol C), and then ribosomal cDNA is depleted. The cDNA fragments are further amplified with primers universal to all libraries. Lastly, the PCR products are purified once more to yield the final cDNA library. Then, the mRNA is fragmented into small pieces using divalent cations under elevated temperature. Then the cleaved RNA fragments werereverse-transcribed to create the final cDNA library in accordance with the protocol for the mRNASeqsample preparation kit (Illumina, San Diego, USA), the average insert size for the paired-endlibraries was 300 bp ( $\pm 50$  bp). And then we performed the paired-end sequencing on an IlluminaHiseq4000 at the (LC Sceiences,USA) following the vendor's recommended protocol.

### 2.3. Quality control and assembly of transcriptome data

Raw data in FASTQ format were first processed through in-house perl scripts. Clean reads were obtained by removing reads containing adapters, reads containing poly-N, and low-quality raw reads. Furthermore, Q20, Q30, and GC contents of the clean data were calculated. All of the downstream sequencing analyses were based on high-quality clean reads. For the RNA-seq data, all clean reads from each sample were aligned to the sheep reference genome ([https://www.ncbi.nlm.nih.gov/genome/83?genome\\_assembly\\_id=351950](https://www.ncbi.nlm.nih.gov/genome/83?genome_assembly_id=351950)) using TopHat v2.0.949. The distribution of known gene types was analyzed by HTSeq software. The mapped reads of each sample were then assembled by both Scripture (beta2) and Cufflinks (v2.1.1) using a reference-based approach.



#### 2.4. RNA-seq reads mapping

We aligned reads to the UCSC (<http://genome.ucsc.edu/>) sheep reference genome using the Tophat package, which initially removes a portion of the reads based on quality information accompanying each read and then maps the reads to the reference genome. Tophat allows multiple alignments to be read (up to 20 by default) and a maximum of two mismatches when mapping the reads to the reference. Tophat builds a database of potential splice junctions and confirms these by comparing the previously unmapped reads against the database of putative junctions.

#### 2.5. Estimation of transcript abundance and differential expression

The mapped reads of each sample were assembled using StringTie. All transcriptomes from the samples were then merged to reconstruct a comprehensive transcriptome using perl scripts. After the final transcriptome was generated, StringTie and Ballgown were used to estimate the expression levels of all transcripts. StringTie was used to determine the expression level of mRNAs by calculating FPKM. Differentially expressed mRNAs and genes were identified if  $\log_2$  (fold change)  $>1$  or  $\log_2$  (fold change)  $<-1$  and by statistical significance (p-value  $<0.05$ ) through the R package Ballgown.

#### 2.6. WGCNA Analysis

The co-expression network analysis was performed using WGCNA (version:1.61) [24]. First, the soft threshold for network construction was selected. The soft threshold constrains the adjacency matrix to assume a continuous value between 0 and 1 so that the constructed network conforms to the power-law distribution and is closer to the real biological network state. Second, the scale-free network was constructed using the blockwiseModules function, followed by the module partition analysis to identify gene co-expression modules, which groups genes with similar patterns of expression. The modules were defined by cutting the clustering tree into branches using a dynamic tree-cutting algorithm and assigning the modules different colors for visualization [24]. The module eigengene (ME) of each module was then calculated. ME represents the expression level for each module. Next, the correlation between ME and the clinical trait in each module was calculated, followed by the determination of gene significance.

#### 2.7. Transcript assembly and identification of candidate lncRNAs

First, Cutadapt was used to remove the reads that were contaminated with adaptors [25], low-quality bases, and undetermined bases. Sequence quality was then verified using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). We used Bowtie2 (FastQC) and topaht2 to map reads to the genome [26]. The mapped reads of each sample were assembled using StringTie [27]. All transcriptomes from the samples were then merged to reconstruct a comprehensive transcriptome using perl scripts. After the transcriptome was generated, StringTie and Ballgown were used to estimate the expression levels of all transcripts. Transcripts that overlapped with known mRNAs and transcripts shorter than 200 bp were discarded. We utilized CPC and CNCI to predict transcripts with coding potential [28]. All transcripts with CPC scores  $<-1$  and CNCI scores  $<0$  were removed. The remaining transcripts were considered lncRNAs.

#### 2.8. Classification of lncRNAs

The annotated lncRNAs were subdivided into four categories based on their locations relative to the nearest protein-coding genes: (i) lncRNAs that do not overlap protein-coding genes (lncRNAs); (ii) lncRNAs located entirely within a protein-coding locus (intragenic lncRNAs); (iii) lncRNAs partially overlapping a protein-coding gene (overlapping lncRNAs); and (iv) lncRNAs that overlapped exons of a protein-coding transcript on the opposite strand (antisense lncRNAs). Perl scripts were developed to classify these four categories.

### 2.9. Quantification and differential expression analysis

The relative abundances of both candidate lncRNAs and coding genes in each sample were computed by calculating the FPKM using Cufflinks (v2.1.1). Differentially expressed lncRNAs in comparison groups were identified using the Cuffdiff program. For biological replicates, transcripts with adjusted P-values <0.05 were considered differentially expressed between the two groups.

### 2.10. Predictions of cis and trans-target genes

To explore the function of lncRNAs, we first predicted the cis and trans-target genes of lncRNAs. We searched for coding, cis-target genes 10k and 100k upstream and downstream, respectively, of candidate lncRNAs and then analyzed their functions. We calculated the expressed correlation between lncRNAs and coding genes with custom scripts and then analyzed their functions through functional enrichment analysis. The trans-target genes and lncRNAs were identified by their expression levels.

### 2.11. SNP analysis

To further characterize the SNPs, we categorized them as genic or intergenic. Approximately half of the SNPs (47 %) were located in genic regions, and the rest was located in intergenic regions. In addition, nonsynonymous SNPs in the exon region were analyzed to determine whether their amino acid character had changed (e.g., hydrophobic to basic or stop codons), given that compositional changes of the amino acids in proteins can result in changes in structural conformation or enzymatic activities and thus generate phenotypic diversity or critical functional variations. First, 20 amino acids were clustered into several character groups. Non-synonymous SNPs that caused amino acid changes from one group to another were searched. Common SNPs in the eight groups representing each line were then classified.

### 2.12. AS data collection

Data on 96 melanoma cases with clinicopathological information were obtained to explore changes in AS events. To analyze the AS profiles for each patient, the SpliceSeq tool (version 2.1), a java application, was used to evaluate the splicing patterns of mRNA in the melanoma cohort. The percent spliced in value was calculated to quantify alternative splicing events and ranged between 0 and 1 for seven types of AS events, including exon skip (ES), alternate terminator (AT), and mutually exclusive (ME).

### 2.13. GO and KEGG enrichment analysis

GO enrichment analysis of the aforementioned groups was conducted using the GSeq R package while correcting for biases in gene length. In addition, KOBAS software and the KEGG database (<http://www.genome.jp/kegg/>) were used to analyze the statistical enrichment of target genes of differentially expressed lncRNAs in KEGG pathways. Lower p-values corresponded to more relevant pathways, and corrected p-values of <0.05 were considered significantly enriched by DEGs.

### 2.14. Statistical analysis

Statistical analysis was performed using SPSS13.0 software. Proportional data were compared using chi-squared analysis or Fisher's exact tests, and differences were considered significant when  $p < 0.05$ . The percentage of blastocyst formed represented the number of blastocysts formed divided by the total number of embryos cultured. The percentage of high-quality blastocyst formed represented the number of high-quality blastocysts divided by the total number of blastocysts cultured.

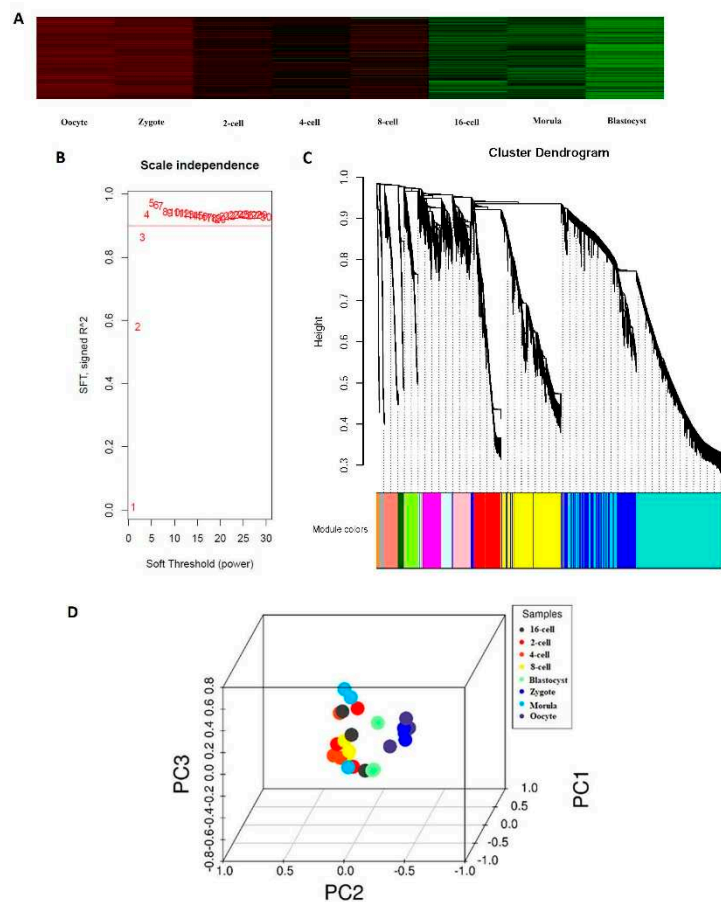
3. Results

3.1. Transcriptome profiles

To establish single-cell transcriptome profiles during the blastomeres of the 2-, 4-, 8-, 16-cell, and morula stage as they underwent asymmetric division, we used previously published protocols; single cells were subjected to RNA-seq sample preparation with several steps modified [8]. Overall, scRNA-seq was performed on 24 cells using the Illumina HiSeq2000 sequencer (table-1). We generated 384 Gb of sequencing data from the 24 cells, with, on average, 10.2 million reads per cell with read lengths of 100 bp. To determine whether these gene expression profiles correlated with developmental stages, we analyzed RNA-seq data from blastocyst cells of the oocytes using a heat map. The greatest changes in gene expression were observed between the morula and blastocyst stages, which may be explained by the major morula-blastocyst transition and are largely representative of the expression patterns occurring during sheep preimplantation development (fig. 1-a). Weighted gene co-expression network analysis (WGCNA) is commonly used to reveal correlations between genes in different samples. Previous WGCNAs have been used to reveal developmentally important genes of the bovine sham nuclear-transfer (shNT) blastocysts [29]. Following the approach of a previous bioinformatics study<sup>24</sup>, the soft threshold for network construction was set to 30 (fig. 1-b). Meanwhile, the fitting degree of the scale-free topological model was 0.85. Thus, this network conformed to the power-law distribution and was closer to the real biological network state. A total of 26 modules (fig. 1-c) were obtained in the current study. The differentially expressed genes (DEGs) in grey were not included in any module; thus, subsequent analyses were no longer performed on DEGs in grey. Mutually exclusive (ME) was in accordance with the expression pattern of DEGs in each module. The pattern of similar within-stage and different between-stage expression patterns were also supported by principal component analysis (PCA) (fig. 1-d). Notably, differences in the expression patterns between oocytes, 4-cell, 8-cell, and blastocyst stages confirmed the key timing of the transcriptome profiles throughout the various stages of sheep development.

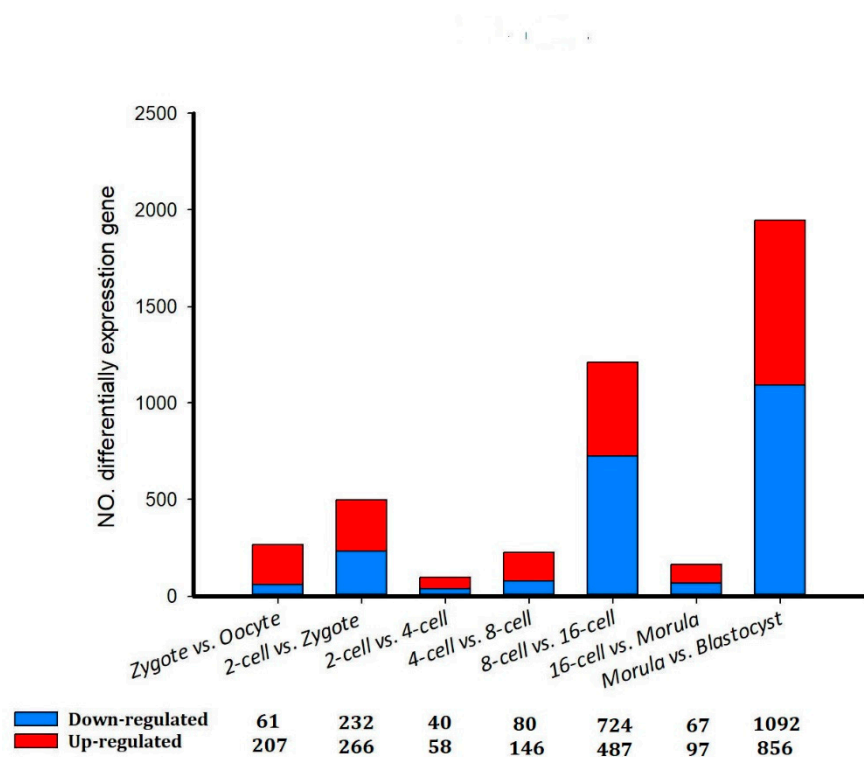
Table 1. Numbers of embryos and cells analyzed by single-cell RNA-Seq analysis.

Sample	No. of sample	No. of single cells
Oocyte	3	3
Zygote	3	3
2-cell	3	3
4-cell	3	3
8-cell	3	3
16-cell	3	3
Morula	3	3
Blastocyst	3	3
Total	24	24



**Figure 1. Expression profiles during sheep development from the oocyte to the preimplantation stage.** (a) Heat map of gene expression profiles correlated with developmental stages. (b) Determination of the soft threshold with the WGCNA algorithm. The approximate scale-free fit index can be attained at the soft-thresholding power of 18. (c) Clustering dendrograms showing 26 modules containing a group of highly connected genes. Each color represents a certain gene module. (d) Principal component analysis. PC1, PC2, and PC3 represent the top three dimensions of the genes showing differential expression among preimplantation blastomeres.





**Figure 2.** Differentially expressed genes in sheep development from the oocyte to the blastocyst stages.

3.2. Differentially expressed genes

Although the total numbers of genes expressed in different stages from the oocyte to the blastocyst varied little, the identities of the genes expressed during early development were dramatically different (Table 2).A total of 2,127 unique genes were identified to be differentially expressed among all of the developmental stages ( $p < 0.05$ ). Consistent with the Pearson correlations between all detected genes, the majority of the DEGs (1,948) were identified between the morula and blastocyst cell stages, indicating that the greatest changes in gene expression occurred during this transition. Among these genes, 1,092 and 856 were down- and up-regulated, respectively (fig.2). For example, between the oocyte and zygote groups, significantly down-regulated genes included *BTF3* (basic transcription factor 3), *TLR1*(toll-like receptor 1) and *SPINT2* (serine peptidase inhibitor, Kunitz type 2), and significantly up-regulated genes included *PEX12* (peroxisomal biogenesis factor 12) and *PGK1* (phosphoglycerate kinase 1). The second-largest change in gene expression occurred between the 8- and 16-cell stages. A total of 1211 DEGs were detected, and 724 and 487 were down- and up-regulated, respectively (fig. 2).

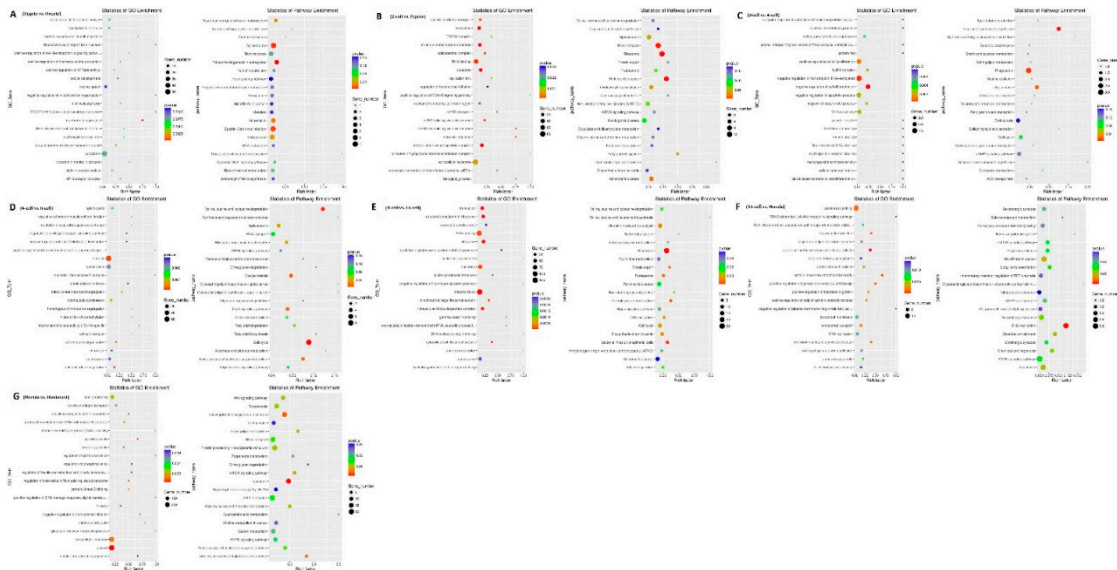
**Table 2.** The numbers of genes detected in sheep from oocyte to blastocyst development.

Stage	No. of genes (FPKM > 0.1)
Oocyte	24729
Zygote	25455
2-cell	26714
4-cell	28948
8-cell	25920
16-cell	24341
Morula	22242
Blastocyst	31157

FPKM: fragments per kilobase of exon per million fragments mapped.

### 3.3. Gene ontology and Kyoto Encyclopedia of Genes and Genomes analyses of DEGs

Between the oocytes and the zygote, 61 and 207 of the 268 DEGs were down- and up-regulated, respectively. These genes represent a rapid degradation of the maternally stored transcripts. Gene ontology (GO) analysis indicated that there was a significant over-representation of elements involved in nuclear speck and cytoplasm. The Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showed that most DEGs were primarily involved in the spliceosome, ribosome biogenesis in eukaryotes, and Epstein-Barr virus infection pathway (Supplemental 1-a, fig. 3-a -1 and -2). Between the zygote and the 2-cell stage, 266 and 232 of the 498 DEGs were down- and up-regulated, respectively. GO analysis indicated that there was a significant over-representation of elements involved in translation, structural constituent of ribosome, RNA binding, and extracellular exosome. The KEGG analysis showed that most DEGs were primarily involved in RNA transport and the ribosome pathway (Supplemental 1-b, fig. 3-b -1 and -2). Between the 2-cell and the 4-cell stage, 58 and 40 of the 98 DEGs were down and up-regulated, respectively. GO analysis indicated that there was a significant over-representation of elements involved in the positive regulation of cell proliferation, negative regulation of transcription, DNA-templated, and negative regulation of cell differentiation. The KEGG analysis showed that most DEGs were primarily involved in the phagosome pathway (Supplemental 1-c, fig. 3-c -1 and -2). Between the 4-cell and the 8-cell stage, 146 and 80 of the 226 DEGs were down- and up-regulated, respectively, which was inconsistent with the findings of a previous study that used scRNA-seq to profile human preimplantation embryos and embryonic stem cells [24,30]. GO analysis indicated that there was a significant over-representation of elements involved in the nucleus. The KEGG analysis showed that most DEGs were primarily involved in the cell cycle pathway (Supplemental 1-d, fig. 3-d -1 and -2). Between the 8-cell and the 16-cell stage, 487 and 724 of the 1211 DEGs were down- and up-regulated, respectively. GO analysis indicated that there was a significant over-representation of elements involved in RNA binding and mitochondrion. The KEGG analysis showed that most of the DEGs were primarily involved in the ribosome pathway (Supplemental 1-e, fig. 3-e -1 and -2). Between the 16-cell and the morula stage, 97 and 67 of the 174 DEGs were down and up-regulated, respectively. GO analysis indicated that there was a significant over-representation of elements involved in transferase activity. The KEGG analysis showed that most of the DEGs were primarily involved in the DNA replication pathway (Supplemental 1-f, fig. 3-f-1 and -2). The majority of the DEGs, 856 and 1092 of the 1948 DEGs were down- and up-regulated, respectively, between the morula and blastocyst stages. GO analysis indicated that there was a significant over-representation of elements involved in extracellular exosome and cytosol. The KEGG analysis showed that most of the DEGs were primarily involved in the transcriptional misregulation in cancer, lysosome, and protein processing in endoplasmic reticulum pathway (Supplemental 1-g, fig. 3-g -1 and -2).



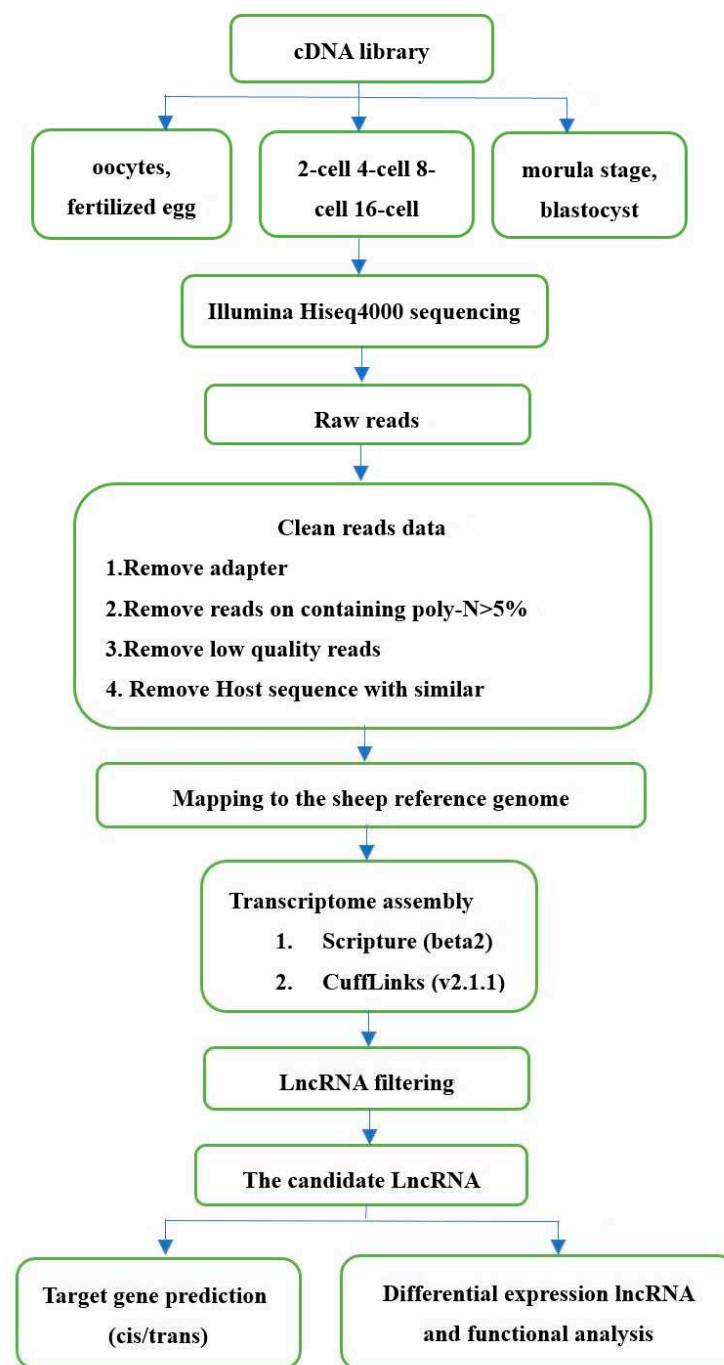
**Figure 3.** Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of differentially expressed genes in sheep development from the oocyte to the blastocyst stages. GO and KEGG analyses between (a) the oocyte and zygote stages, (b) the zygote and 2-cell stages, (c) the 2-cell and 4-cell stages, (d) the 4-cell and 8-cell stages, (e) the 8-cell and 16-cell stages, (f) the 16-cell and morula stages, and (g) the morula and blastocyst stages.

3.4. Genome-wide discovery and identification of lncRNAs

To identify new lncRNAs involved in sheep development from the oocyte to the blastocyst stages, cell samples from oocytes, fertilized eggs, and blastomeres of the 2-, 4-, 8-, 16-cell, morula, and blastocyst stages were collected. Transcriptome sequencing was then performed using the Illumina HiSeqTM 4000 platform. An overview of the analysis pipeline is shown in Figure 4. After removing adaptor reads, reads containing poly-N>10%, and low-quality reads, clean reads were obtained for each sample and used in the following analyses. The GC content of each sample was between 43.5 and 45%. Thus, approximately 75.28–91.45% of the total clean reads could be mapped to the mouse reference genome sequence using TopHat v2.0.9 (Table 3). The different gene subtypes of the above-mapped reads are shown in Figure 1B and are based on genomic overlap with existing annotations using the HTseq program. A total of 274,470 assembled transcripts were produced using both Scripture (beta2) and Cufflinks (v2.1.1).

**Table 3.** Summary of read filter and alignment.

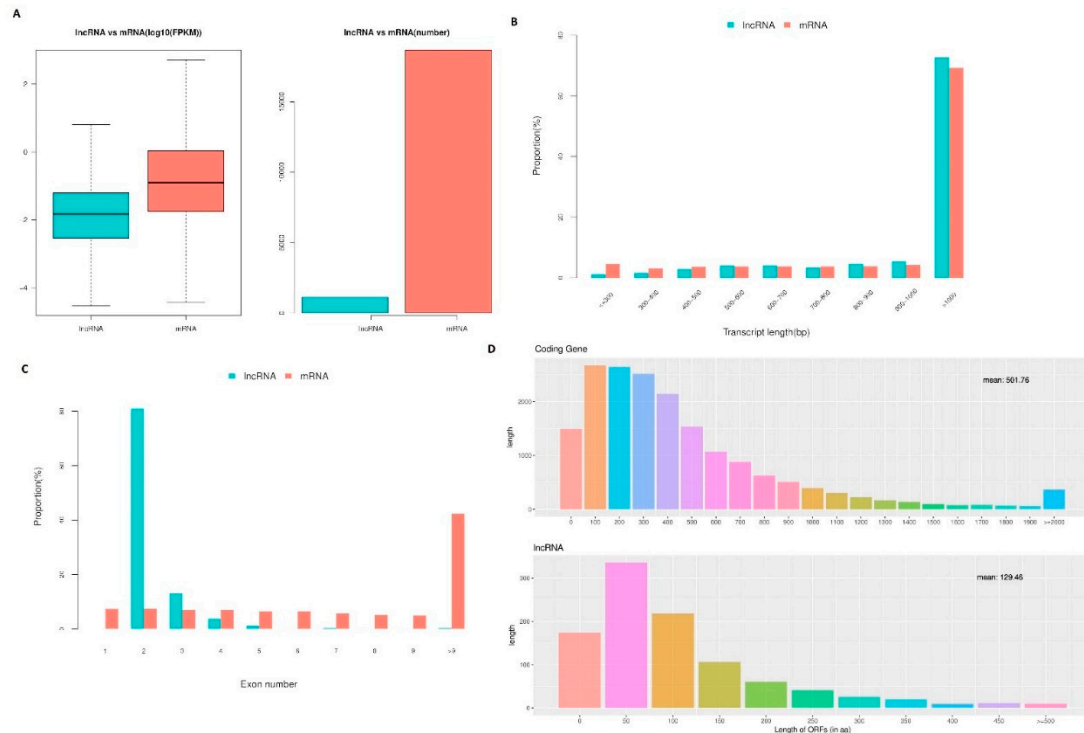
Sample	Raw reads	Clean reads	Clean bases	Error rate (%)	GC content (%)
oocyte	102,383,290	86,741,418	13.01G	0.04	43.50
fertilized egg	111,921,426	87,081,144	13.06G	0.02	44.50
2-cell	123,273,652	95,679,634	14.35G	0.04	44
4-cell	108,522,186	96,419,102	14.46G	0.03	44
8-cell	142,178,172	123,561,104	18.53G	0.02	45
16-cell	133,023,048	90,477,914	13.57G	0.04	43.5
morula	122,556,646	90,506,796	13.58G	0.02	44.5
blastula	141,488,370	105,354,406	15.80G	0.03	44



**Figure 4.** Pipeline for identification of lncRNAs.

### 3.5. Features of lncRNAs

The fragments per kilobase of exon per million fragments mapped (FPKM) values and numbers of lncRNAs demonstrated that lncRNAs in cell samples from the oocyte, fertilized egg, 2-, 4-, 8-, 16-cell, morula, and blastocyst stages were expressed at lower levels compared with the levels of protein-coding genes expressed (fig. 5-a). However, the lncRNA transcript length was mostly  $\geq 1000$  bp, which was not significantly different compared with transcript lengths observed for protein-coding genes (fig. 5-b). In addition, significant differences in the distributions of exon number between protein-coding genes and lncRNAs were also detected, and 82.41% of all lncRNAs only contained two exons (fig. 5-c). Furthermore, most of the lncRNAs contained relatively shorter open reading frames (ORFs) compared with protein-coding genes (fig. 5-d).



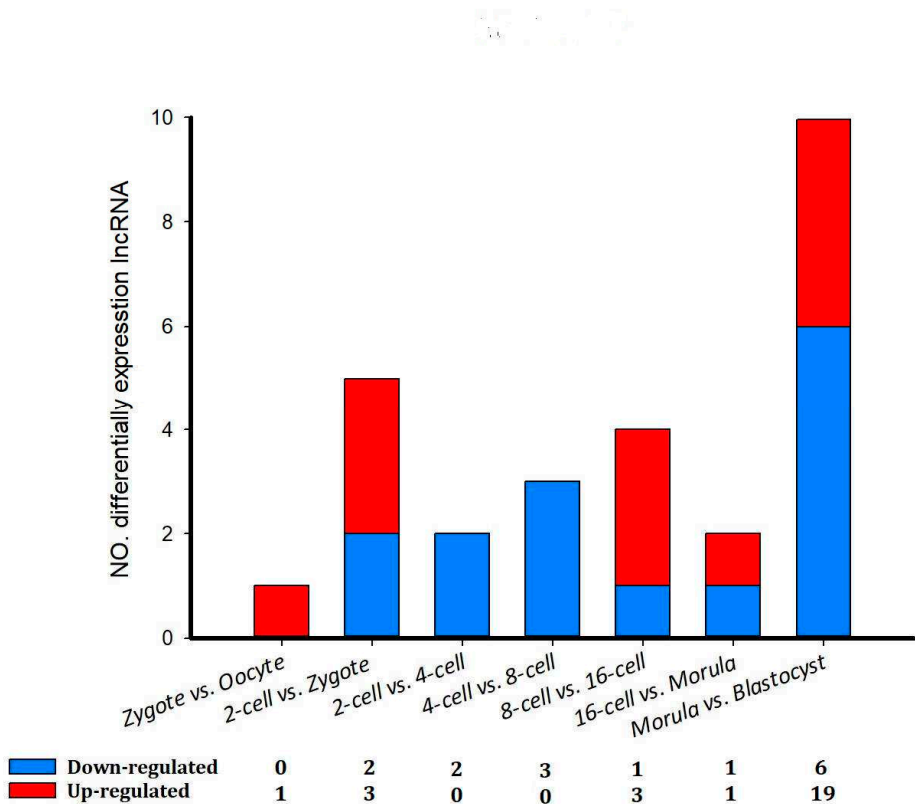
**Figure 5. LncRNA characteristics in sheep development from the oocyte to the blastocyst stages.**

(a) Comparison of the expression level between lncRNA and protein-coding genes in terms of fragments per kilobase of exon per million fragments mapped (FPKM). The FPKM distribution of lncRNAs in mouse uterus was lower than that of protein-coding genes. (b) Distribution of transcript lengths in the lncRNAs and protein-coding genes. The average size of lncRNA transcripts was generally shorter than that of protein-coding genes. (c) The number of exons in lncRNAs and protein-coding genes. A total of 88.41% of the lncRNAs contained two to four exons, while the majority of protein-coding genes contained more than 10 exons. (d) The number of ORFs identified in the lncRNAs and protein-coding genes using Estscan. As expected, the ORFs of lncRNAs were substantially shorter than the ORFs in protein-coding genes.

### 3.6. Differentially expressed lncRNAs

A total of 42 differentially expressed lncRNAs ( $p < 0.05$ ; 52 transcript isoforms) were detected (fig. 6). The most noticeable changes in lncRNA expression occurred between the morula and blastocyst stages, in which 19 (23 transcript isoforms) were significantly up-regulated and six (10 transcript isoforms) were down-regulated (fig. 6). Overall, differentially expressed lncRNA transcripts were fewer in number in sheep compared with the implantation and inter-implantation sites of day-5 pregnant mice [31].





**Figure 6.** The number of differentially expressed lncRNAs in eight comparison groups in sheep development from the oocyte to the blastocyst stages.

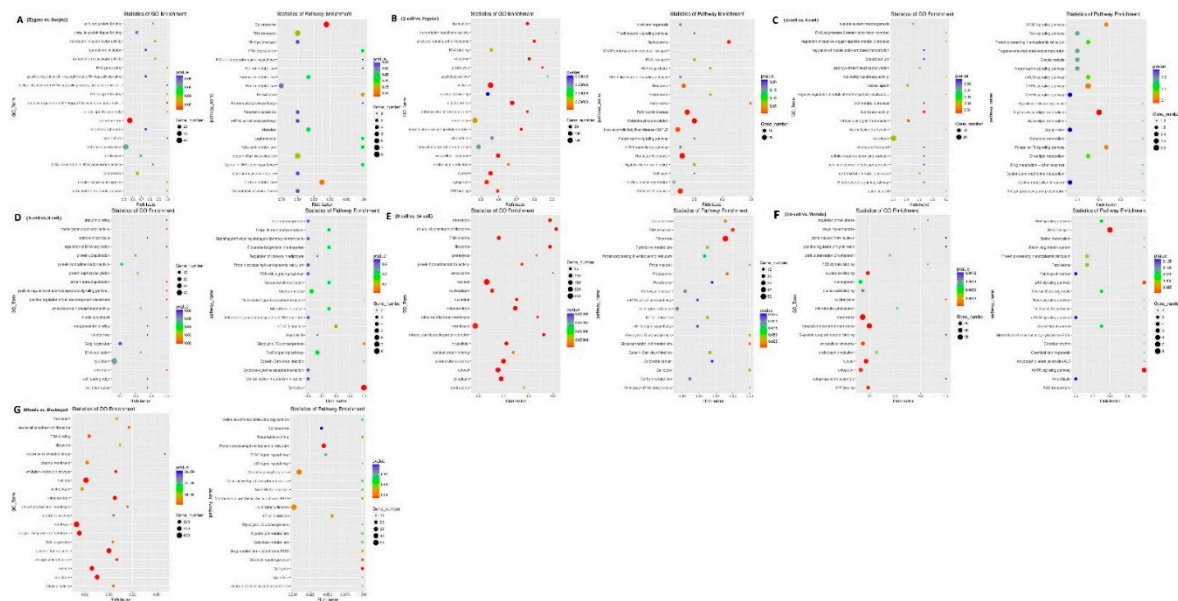
3.7. *cis* and *trans*-target genes of lncRNAs

To investigate the function of lncRNAs, we first predicted the putative lncRNA *cis*- and *trans*-regulatory target genes. Protein-coding genes located within 100 kb upstream and downstream of lncRNAs were considered *cis*-targets (Supplemental 2-a). The *trans*-target genes located far from lncRNAs are shown in Supplemental 2-b.

3.8. *Functional analysis of differentially expressed lncRNAs*

To predict the function of lncRNAs during the different aforementioned stages, GO and KEGG analyses of the *cis* and *trans*-target genes of the lncRNAs in the eight comparison groups were performed. GO analysis of the *cis*-targets revealed only one significantly enriched GO term (mitochondrion) in the oocyte vs. zygote stages. The KEGG analysis revealed that the significantly enriched pathways in the oocyte vs. zygote stages were “Spliceosome” and “Carbon metabolism” (corrected p-value <0.05, fig. 7-a-1 and-2 and Supplemental 3-1). The significantly enriched GO terms of the *cis*-targets in the zygote vs. 2-cell stages, which represented biological processes and molecular functions, were associated primarily with nucleus, mitochondrion, extracellular exosome, cytosol, and cytoplasm. The KEGG analysis revealed that the significantly enriched *cis* pathways in the oocyte vs. zygote stages were “Non-alcoholic fatty liver disease,” “Huntington’s disease,” and “Alzheimer’s disease” (corrected p-value <0.05, fig. 7-b-1and-2 and Supplemental 3-2). There was only one significantly enriched GO term detected in the 2-cell vs. 4-cell stages: intracellular signal transduction. The KEGG analysis revealed that the significantly enriched *cis* pathways were the “MAPK signaling pathway” and “Glycerophospholipid metabolism” (corrected p-value <0.05, fig. 7-c and Supplemental 3-3). Furthermore, no GO terms were significantly enriched in the 4-cell vs. 8-cell stages. The KEGG analysis revealed that the only significantly enriched *cis* pathway was “cell cycle” (corrected p-value <0.05, fig. 7-c and Supplemental 3-4). In the 8-cell vs. 16-cell stages, there were

additional significantly enriched GO terms: nucleus, membrane, and others. The KEGG analysis revealed that the only significantly enriched cis pathway was “Ribosome” (corrected p-value <0.05, fig. 7-e and Supplemental 3-5). Seven significantly enriched GO terms were detected in the 16-cell vs. morula stages: nucleotide binding, membrane, and others. The KEGG analysis revealed that the significantly enriched cis pathways were “RNA transport” and “AMPK signaling pathway” (corrected p-value <0.05, fig. 4-f and Supplemental 3-6). In the morula vs. blastula stages, there were additional significantly enriched GO terms: RNA binding, nucleus, and others. The KEGG analysis revealed that the significantly enriched cis pathways were “Huntington’s disease” and “Oxidative phosphorylation” (corrected p-value <0.05, fig. 7-g and Supplemental 3-7).



**Figure 7. GO and KEGG analysis with the cis and trans-target genes of lncRNAs.** GO and KEGG analyses between (a) oocyte and zygote stages, (b) zygote and 2-cell stages, (c) 2-cell and 4-cell stages, (d) 4-cell and 8-cell stages, (e) 8-cell and 16-cell stages, (f) 16-cell and morula stages, and (g) morula and blastocyst stages.

### 3.9. Distribution of different SNP and indel types during sheep from oocyte to blastocyst development

Single-nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans and drive phenotypic variation. Due to evolutionary conservation, SNPs and indels (insertion and deletions) are depleted in functionally important sequence elements [32,33]. Here, using the SAMtools/Popoolation software package. A total of 4,352,847 putative SNPs and 297,411 INDEL were predicted, with an average of 181,368 SNPs and 12,392 INDEL per sample, respectively. After removing redundant SNPs among all samples, we had 678,433 and 8,454 INDEL from SAMtools/Popoolation2 (Supplemental 4). Then, concerning all the putative SNPs during sheep from oocyte to pre-implantation development, there were 79516, 70858, 99129, 115926, 72615, 78782, 66420, 224557 intergenic SNPs. In addition, 47852, 40178, 60970, 67126, 37762, 64925, 39054, 201642 SNPs from different stages during sheep from oocyte to pre-implantation development were genic, and defined exactly as in the 5'UTR, 3'UTR, upstream and downstream of protein-coding genes. (Table 6). Furthermore, in these three SNP datasets, there were large percentages of intergenic (including upstream/downstream) SNPs (37–49%). there were 79516, 70858, 99129, 115926, 72615, 78782, 66420, 224557 intergenic SNPs. In addition, 47852, 40178, 60970, 67126, 37762, 64925, 39054, 201642 SNPs from different stages during sheep from oocyte to pre-implantation development were genic, and of these genic SNPs (Table 4).

Table 4. Summary of SNP and indel classification for different sets.

Functional Class	SNP								INDEL							
	Oocyte	Zygote	2-cell	4-cell	8-cell	16-cell	Morula	Blastocyst	Oocyte	Zygote	2-cell	4-cell	8-cell	16-cell	Morula	Blastocyst
Intergenic	79516	70858	99129	115926	72615	78782	66420	224557	5650	5260	7088	8129	6163	4939	4765	15130
Intronic	47852	40178	60970	67126	37762	64925	39054	201642	3104	2831	3973	4172	2793	3580	2566	15373
Exonic	28898	24940	33951	35201	26715	32414	27569	66325	336	298	438	411	415	326	440	512
3'UTR	3991	4763	5879	5761	4773	4822	3991	8356	502	564	666	651	598	547	577	862
5'UTR	1075	904	1142	1269	964	1265	1075	2145	34	32	35	42	35	41	44	60
upstream	4343	3612	5330	5836	3937	5551	4343	9568	243	233	359	408	285	306	299	588
downstream	14572	16348	19799	20220	16904	16251	14572	27967	1845	1814	2268	2209	2072	1749	2010	2886
splicing	632	375	421	494	390	390	632	525	47	43	53	52	47	44	53	71
Total number	184380	162027	226690	251891	164117	204485	157743	541244	11763	11076	14882	16076	12408	11534	10755	35487

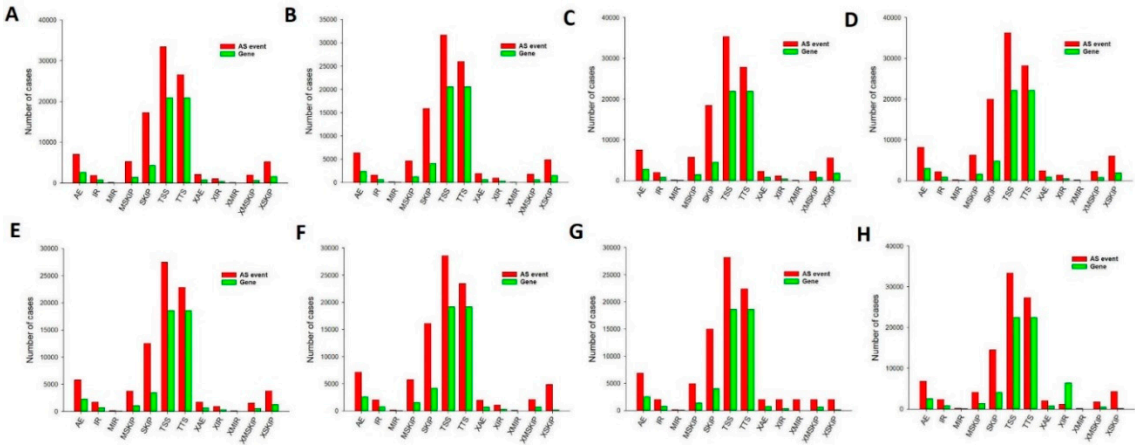
3.10. SNP and indel functional annotation

Functional annotation of SNPs with allelic imbalances was performed using the Blast2GO suite [33–36]. The SNP-flanking sequences were searched against the NCBI nr-protein database using BLASTx. Associated genes and GO terms were then obtained. In the biological processes’ category, SNP genes were associated with various cellular processes that were primarily involved in development-related mechanisms, including the regulation of the MyD88-dependent toll-like receptor signaling pathway, regulation of metabolic and oxidation-reduction processes, and protein translation. In the molecular function category, SNP-containing genes were associated with binding phosphoprotein, nucleic acid, and actin. In addition, a significant number of the genes were associated with transferase, motor, oxidoreductase, and structural molecule activities. In the cellular component category, many of the genes were associated with the cytoplasmic compartment, membranes, myosin complex, and extracellular region compartment (Supplemental 4 and 5), which is consistent with the findings of previous studies [32].

Additionally, KEGG pathway mapping was used to assign functions to the SNP-containing transcripts. A search of transcripts against the KEGG database yielded 1043 transcripts (13.15%) with significant hits to 632 KEGG Orthologies belonging to different pathways (Supplemental 4 and 5).

3.11. Overview of AS events during sheep from oocyte to blastocyst development

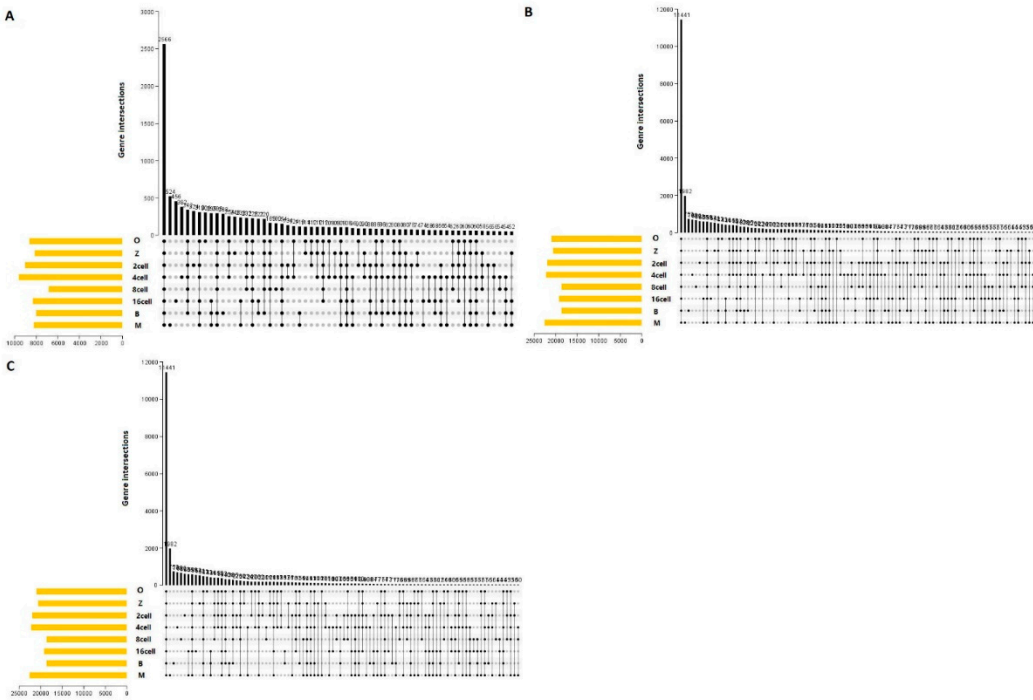
Splicing events were comprehensively analyzed for 24 single cells based on relevant RNA-seq data. In oocyte stage, there were totally 101835 AS event were detected in 54266 genes, comprising 7003 alternative exon ends (AE) events were detected in 2603 genes, 1862 retention of single (IR) events in 723 genes, 146 multiple (MIR) intron events in 56 genes, 5260 multi-exon SKIP (MSKIP) events in 1400 genes, 17236 Skipped exon (SKIP) events in 4310 genes, 33409 Alternative 5' first exon (TSS) events in 20899 genes, 26563 alternative transcription termination site (TTS) events in 20900 genes, 2063 approximate AE (XAE) events in 729 genes, 1055 approximate IR (XIR) events in 359 genes, 76 approximate MIR (XMIR) events in 31 genes, 1969 approximate MSKIP (XMSKIP) events in 631 genes, and 5193 approximate SKIP (XSKIP) events in 1625 genes (fig8-A). Next, Splicing events comprehensively analyzed in fertilized egg stage, 2-cell, 4-cell, 8cell, 16 cell, morula and blastocyst shown in fig8(B-H).Of those stags, in morula and blastula, only one type of AS event was detected in most genes, although there were some exceptions; generally, it was demonstrated that 3 or more splicing events could be attributed to one gene, with a maximum of 5 types of AS events observable for a single gene. However, TSS was the predominant type of event in all the histologic STS subtypes, which revealed that TSS was the most common splicing event during sheep from oocyte to blastocyst development.



**Figure 8.** The number of alternative splicing events and associated genes in sheep development from the oocyte to the blastocyst stages. TSS was the most frequent of the eleven types of events. (a) Oocyte stage. (b) Zygote stage. (c) 2-cell stage. (d) 4-cell stage. (e) 8-cell stage. (f) 16-cell stage. (g) Morula stage. (h) Blastocyst stage. AE alternative exon ends, AD alternate donor, AP alternate promoter, AT alternate terminator, ES exon skip, ME mutually exclusive exon, RI retained intron.

3.13. Associated AS events

The main AS events were SKIP, TSS, and TTS (fig. 1). Therefore, we analyzed associated AS events and genes by UpSet plot. SKIP was detected in 2566 genes at every stage (fig. 9-a). Furthermore, TSS and TTS were detected in 11441 genes at every stage (fig. 9-b and c). More than one AS event could occur in one gene, and up to three types of splicing events were identified in one gene. Thus, one gene might have two or more AS events that were markedly related to the PFI of PTC patients. The UpSet plot revealed that TSS was the most common prognosis-related event.



**Figure 9.** UpSet plot of the three main alternative splicing events at different stages in sheep development from the oocyte to the blastocyst stages. UpSet plot of the interactions of the alternative splicing events associated with genes. (a) UpSet plot of the interactions of SKIP. (b) UpSet plot of the interactions of TSS. (c) UpSet plot of the interactions of TTS. TTS alternative transcription termination site.



### 3.14. Molecular characteristics of the most important AS events

Based on UpSet plot, there were 13663 genes with one or more AS events at different stages (fig. 10-a). We then carried out GO and KEGG analysis. The functional annotations revealed that “regulation of transcription, DNA templated (884 genes),” “transport (609 genes),” and “proteolysis (397 genes)” were the three most significant biological process terms. “Membrane (4323 genes),” “Nucleus (3966 genes),” and “integral component of membrane (3055 genes)” were the three most significant cellular component terms. For molecular function, “ATP binding (1498 genes),” “nucleotide binding (1441 genes),” and “nucleic acid binding (1412 genes)” were the three most enriched categories (fig. 10-b).

Furthermore, we found that related pathways were metabolism, environmental information processing, and human diseases. The “metabolism” and human diseases correlated pathways were mostly genes involved in purine metabolism (131 genes), pyrimidine metabolism (93 genes), and pathways in cancer (181 genes) (fig. 10-c). In our study, there were 13663 genes with one more AS event suggesting that AS may play an important role at different developmental stages. Furthermore, “ATP binding (1498 genes),” “nucleotide binding (1441 genes),” and “nucleic acid binding (1412 genes)” were the three most enriched categories, suggesting that alternative protein isoforms with distinct functions are expressed. Thus, defects in splicing control might be able to induce losses in function and severe phenotypes and require further study.

## 4. Discussion

In recent years, much research has focused on the study of the development of blastomeres of the 4- and 8-cell embryos of mice [5], cattle [6], and goats [7]. However, the exact mechanism and the developmental patterns underlying how the blastomeres of the 2-, 4-, 8-, 16-cell, morula, and blastocyst stages undergo asymmetric division are still unclear. For the first time, we used scRNA-seq to study the transcriptome profiles during sheep development from the oocyte to the blastocyst stages. Our data showed that from the 4-cell to the 8-cell stage, there were no noticeable changes in the transcriptome profile as has been shown for the 4- and 8-cell embryos of mice [37,6,7] (fig. 1-a). However, the first major change was noted between the 8-cell and 16-cell stages, which is similar to the pattern observed in the bovine embryonic genome. The greatest changes in gene expression were observed between the morula and blastocyst stages, which may be explained by a major morula-blastocyst transition that results in expression patterns characteristic of sheep preimplantation development [11,38]. Furthermore, 1,092 and 856 DEGs were down- and up-regulated, respectively (fig. 2). *BTF3* (basic transcription factor 3), *TLR1* (toll-like receptor 1) and *SPINT2* (serine peptidase inhibitor, Kunitz type 2) were markedly up-regulated, while *PEX12* (peroxisomal biogenesis factor 12) and *PGK1* (phosphoglycerate kinase 1) were significantly down-regulated between the oocyte and zygote stages. In others reports, BTF3 as one of importance transcription factor was improve in gastric cancer development and progression by enhance transcription [39]. TLR4 enhances blastocyst attachment to endometrial cells in mice via miR-Let-7a suggested that inflammatory responses are beneficial in the fetomaternal interface and supposedly accelerate the chances for successful implantation. In our study, TLR1 was markedly up expression from oocyte to blastocyst development [40]. Therefore, the role of inflammatory responses is interesting need further study. GO and KEGG analyses of DEGs were then conducted. At different stages, GO analysis indicated that there was a significant over-representation of elements involved in nuclear speck and cytoplasm. The KEGG analysis revealed that most of the DEGs were primarily involved in the spliceosome, ribosome biogenesis in eukaryotes, and the Epstein-Barr virus infection pathway, RNA transport, ribosome pathway, transferase activity, and others. The above pathways have also been reported in the development of bovine and monkey embryos [11,41,42]. Therefore, the transcriptome profiles during sheep from oocyte to blastocyst development were shown the same pattern.

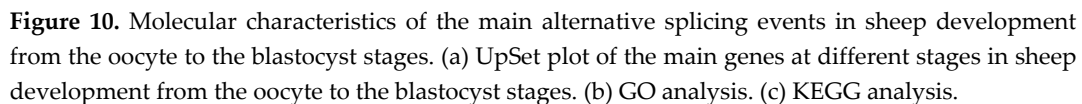
We then characterized the features of the lncRNAs and their target genes. The lncRNAs in cell samples from the oocyte, fertilized egg, 2-, 4-, 8-, 16-cell, morula, and blastocyst stages had lower expression levels compared with the expression levels observed in protein-coding genes. However, the transcript lengths of lncRNAs were mostly  $\geq 1000$  bp, and the mean transcript length of lncRNAs



was not significantly different relative to the mean transcript length of protein-coding genes (fig. 5-b). In addition, there were a total of 42 differentially expressed lncRNAs (52 transcript isoforms) (fig. 6). The major changes in lncRNA expression occurred between the morula and blastocyst stages, of which 19 (23 transcript isoforms) lncRNAs were significantly up-regulated, such as lnc MSTRG.2676, lnc MSTRG.3585, and six (10 transcript isoforms) lncRNAs were down-regulated, such as lnc MSTRG.8262 and lncMSTRG.11966 (fig. 6). Overall, there were fewer differentially expressed lncRNA transcripts during sheep development compared with the number of differentially expressed transcripts detected in implantation and inter-implantations sites in day-5 pregnant mice [31]. GO and KEGG analyses of the cis and trans-target genes of the lncRNAs in the eight comparison groups were performed. GO analysis of the cis lncRNA targets revealed the following significantly enriched GO terms: RNA binding, nucleus, and others. The KEGG analysis revealed that the significantly enriched cis pathways were "Huntington's disease" and "Oxidative phosphorylation" between the morula and blastula groups (corrected p-value <0.05, fig. 7-g and Supplemental 3-7).

We then studied the distribution of different SNP and indel types (Table 4). Approximately 10% intergenic and 30% non-coding SNPs have been reported in humans from RNA-seq data [43]. The high percentages of intergenic SNPs in humans may be partially explained by the incomplete annotation of protein-coding genes and exons in the current version of the rainbow trout reference genome sequence [36]. We then conducted GO analysis of the associated genes. In the biological process's category, SNP genes were associated with various cellular processes that were primarily involved in development-related mechanisms, including regulation of the MyD88-dependent toll-like receptor signaling pathway, regulation of metabolic and oxidation-reduction processes, and protein translation. In the molecular function category, SNP-containing genes were associated with binding phosphoprotein, nucleic acid, and actin. In addition, many genes were associated with transferase, motor, oxidoreductase, and structural molecule activities. In the cellular component category, many of the genes were associated with the cytoplasmic compartment, membranes, myosin complex, and extracellular region compartment (Supplemental 4 and 5), which is consistent with the findings of previous studies [32].

Finally, we also detected AS events. There are relatively few studies that have reported the numbers of AS events during development from the oocyte to the blastocyst. In this study, splicing events were comprehensively analyzed for 24 single cells based on relevant RNA-seq data. For example, there were a total of 101835 AS events that were detected in 54266 genes, comprising 7003 AE events detected in 2603 genes in the oocyte stage (fig. 8a). AS can control the transcriptional identity of the maternal transcriptome by the RNA-binding protein, which is essential for the development of fertilized-competent oocytes [44]. Therefore, our comprehensive analysis of AS suggests that AS plays an important role in sheep development. GO and KEGG analyses showed that there were 13663 genes with one or more AS events at different stages during sheep development based on the UpSet plot (fig. 10-a). The most significant biological process terms were "regulation of transcription, DNA templated," "transport," and "proteolysis." The three most significant cellular component terms were "membrane," "nucleus," and "integral component of membrane." The significant terms for molecular function were "ATP binding," "nucleotide binding," and "nucleic acid binding" (fig. 10-b). In a previous study, AS has been shown to play an important role in the protein-coding genes of mouse oocytes and zygotes by RNA-binding protein, and the correct combination of exons through AS ensures that gene isoforms are expressed for the specific context in which they are required [45,46]. Furthermore, the "metabolism," "Genetic information processing" and "human diseases" correlated pathways were mostly genes involved in purine metabolism, RNA transport and pathways in cancer (fig. 10-c). In metabolomic analyses of fetal germ cells in mice on embryonic day (E)13.5 and E18.5, purine metabolism was involved in demonstrate sex- and developmental stage-dependent changes in these processes [47]. RNA transport in our study were mostly genes involved which also as one of importance AS event is frequent across developmental stages and tissues in embryonic day 8.5, 9.5 and 11.5 mouse embryos and placenta [48]. Furthermore, We detected 13663 genes with one or more AS events, suggesting that AS may play an important role in sheep development. Furthermore, "ATP binding (1498 genes)," "nucleotide binding (1441 genes),"



Here, we conducted a scRNA-seq survey of cells from sheep from the oocyte to the blastocyst developmental stages. We then carried out experiments to identify and functionally analyze the transcriptome profiles, long non-coding RNAs, single-nucleotide polymorphisms (SNPs), and alternative splicing (AS). We founded that between the oocyte and zygote groups significantly down-regulated genes and the second-largest change in gene expression occurred between the 8- and 16-cell stages. We used various methods to construct a profile to characterize cellular and molecular features and systematically analyze the related GO and KEGG profile of cells of all stages from the oocyte to the blastocyst. This large-scale, single-cell atlas provides key cellular information and will likely assist clinical studies in improving preimplantation genetic diagnosis.

**Funding:** This work was supported for the Henan Beef Cattle Industrial Technology System (HARS-22-13-S), The National Beef Cattle Industrial Technology System (No. CARS-37), Scientific and Technological Key Project of Henan Province (222102110069, 222102110018, 212102110008), the Science-Technology Foundation for Outstanding Young Scientists of Henan Academy of Agricultural Sciences (2022YQ20), the Key Scientific and Technological Special Projects of Henan Province (201300111200), funds for Science Research and Development of Henan Academy of Agricultural Science (221100110200), the Breeding and production of cattle and sheep by scientific and technological innovation team of Henan Academy of Agricultural Sciences (2023TD28).

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References

1. Zhang K: Huang K, Luo Y1, LiS. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics*. 2014;15:845.
2. Daughtry BL, Rosenkrantz JL, Lazar NH, Fei SS, Redmayne N, Torkenczy KA, Adey A, Yan M, Gao L, Park B, Nevenon KA, Carbone L, Chavez SL. Single-cell sequencing of primate preimplantation embryos reveals chromosome elimination via cellular fragmentation and blastomere exclusion. *Genome Res*. 2019;29(3):367-382.
3. Cheung VG, Spielman RS. The genetics of variation in gene expression. *Nat Genet*. 2002; 32: S522–S525.
4. Cheung VG, Bruzel A, Burdick JT, Morley M, Devlin JL, Spielman RS. Monozygotic twins reveal germline contribution to allelic expression differences. *Am J Hum Genet*. 2008;82(6):1357-60.
5. Zhang X, Li T, Zhang L, Jiang L, Cui T, Yuan X, Wang C, Liu Z, Zhang Y, Li W, Zhou Q. Individual blastomeres of 4- and 8-cell embryos have ability to develop into a full organism in mouse. *J Genet Genomics*. 2018;45(12):677-680.
6. Daigneault BW, Rajput S, Smith GW, Ross PJ. Required for Expanded Bovine Blastocyst Formation. *Sci Rep*. 2018;8(1):7753.
7. Deng M, Liu Z, Ren C, Zhang G, Pang J, Zhang Y, Wang F, Wan Y. Long noncoding RNAs exchange during zygotic genome activation in goat. *Biol Reprod*. 2018;99(4):707-717.
8. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol*. 2013;20(9):1131-9.
9. Blakeley P, Fogarty NM, del Valle I, Wamaitha SE, Hu TX, Elder K, Snell P, Christie L, Robson P, Niakan KK. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*. 2015;142(18):3151-65.
10. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. *Nat Methods*. 2011;8(4):S6–S11.
11. Liu XM, Wang YK, Liu YH, Yu XX, Wang PC, Li X, Du ZQ, Yang CX. Single-cell transcriptome sequencing reveals that cell division cycle 5-like protein is essential for porcine oocyte maturation. *J Biol Chem*. 2018;293(5):1767-1780.
12. Yin XY, Cheng GH, Guo HY, Wang Q, Li YJ, Zhang H. Single cell transcriptome profiling revealed differences in gene expression during oocyte maturation in Haimen white goats. *Genet Mol Res*. 2017;16(1).
13. Petr Svoboda. Long and small noncoding RNAs during oocyte-to-embryo transition in mammals. *Biochem Soc Trans*. 2017;45(5):1117-1124.
14. Abbastabar M, Sarfi M, Golestani A, Khalili E. lncRNA involvement in hepatocellular carcinoma metastasis and prognosis. *EXCLI J*. 2018;17:900-913.
15. Youness RA, Gad MZ. Long non-coding RNAs: Functional regulatory players in breast cancer. *Noncoding RNA Res*. 2019;4(1):36-44.
16. Svoboda P. Long and small noncoding RNAs during oocyte-to-embryo transition in mammals. *Biochem Soc Trans*. 2017;45(5):1117-1124.
17. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25(18):1915-27.
18. Rivera RM, Ross JW. Epigenetics in fertilization and preimplantation embryo development. *Prog Biophys Mol Biol*. 2013;113(3):423–432.
19. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res*. 2015;43(8):4363-4.
20. Veselovska L, Smallwood SA, Saadeh H, Stewart KR, Krueger F, Maupetit-Méhouas S, Arnaud P, Tomizawa S, Andrews S, Kelsey G. Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. *Genome Biol*. 2015;16:209.
21. Wu F, Liu Y, Wu Q, Li D, Zhang L, Wu X, Wang R, Zhang D, Gao S, Li W. Long non-coding RNAs potentially function synergistically in the cellular reprogramming of SCNT embryos. *BMC Genomics*. 2018;19(1):631.
22. Karlic R, Ganesh S, Franke V, Svobodova E, Urbanova J, Suzuki Y, Aoki F, Vlahovicek K, Svoboda P. Long non-coding RNA exchange during the oocyte-to-embryo transition in mice. *DNA Res*. 2017;24(2):219-220.
23. Caballero J, Gilbert I, Fournier E, Gagné D, Scantland S, Macaulay A, Robert C. Exploring the function of long non-coding RNA in the development of bovine early embryos. *Reprod Fertil Dev*. 2014;27(1):40-52.

24. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
25. Martin, M.. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 2011,17(1), pp-10
26. Langmead, B., & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 2012,9(4):357-359.
27. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 2015,33(3):290-295.
28. Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., ... & Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, 2013. gkt646.
29. Min B, Park JS, Kang YK. Determination of Oocyte-Manipulation, Zygote-Manipulation, and Genome-Reprogramming Effects on the Transcriptomes of Bovine Blastocysts. *Front Genet*. 2018;9:143.
30. Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res*. 2010;20(6):804-15.
31. Wang Qi, Wang Nan, Cai Rui, Zhao Fan, Xiong Yongjie, Li Xiao, Wang Aihua, Lin Pengfei and Jin Yaping. Genome-wide analysis and functional prediction of long non-coding RNAs in mouse uterus during the implantation window. *Oncotarget*, 2017, 8(48): 84360-84372.
32. Al-Tobasei R, Ali A, Leeds TD, Liu S, Palti Y, Kenney B, Salem M. Identification of SNPs associated with muscle yield and quality traits using allelic-imbalance analyses of pooled RNA-Seq samples in rainbow trout. *BMC Genomics*. 2017,18(1):582.
33. Neininger K, Marschall T, Helms V. SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. *PLoS One*. 2019,14(4):e0214816.
34. Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. SNP calling by sequencing pooled samples. *BMC Bioinformatics*. 2012,13:239
35. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013,93(4):641–51.
36. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun*. 2014,5:3657
37. Petropoulos S, Edsgård D, Reinius B, Deng Q, Panula SP, Codeluppi S, Plaza Reyes A, Linnarsson S, Sandberg R, Lanner F. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*. 2016,165(4):1012-26.
38. Macaulay IC, Voet T. Single cell genomics: advances and future perspectives. *PLoS Genet*. 2014,10(1): e1004126.
39. Qi Liu, Jian-Ping Zhou, Bin Li, Zhong-Cheng Huang, Hong-Yu Dong, Guang-Yi Li, Ke Zhou, and Shao-Lin Nie. Basic transcription factor 3 is involved in gastric cancer development and progression. *World J Gastroenterol*. 2013,19(28): 4495–4503.
40. Sara Hosseini, Samaneh Hosseini, Mohammad Salehi. Upregulation of Toll-like receptor 4 through anti-miR-Let-7a enhances blastocyst attachment to endometrial cells in mice. *J Cell Physiol*. 2020,235(12):9752-9762.
41. Lavagi I, Krebs S, Simmet K, Beck A, Zakhartchenko V, Wolf E, Blum H. Single-cell RNA sequencing reveals developmental heterogeneity of blastomeres during major genome activation in bovine embryos. *Sci Rep*. 2018,8(1):4071.
42. Razza EM, Sudano MJ, Fontes PK, Franchi FF, Belaz KRA, Santos PH, Castilho ACS5, Rocha DFO, Eberlin MN, Machado MF, Nogueira MFG. Treatment with cyclic adenosine monophosphate modulators prior to in vitro maturation alters the lipid composition and transcript profile of bovine cumulus-oocyte complexes and blastocysts. *Reprod Fertil Dev*. 2018,30(10):1314-1328.
43. Zhang Feng, James R Lupski. Non-coding genetic variants in human disease. *Hum Mol Genet*. 2015,24(R1): R102-10.
44. Do DV, Strauss B, Cukuroglu E, Macaulay I, Wee KB, Hu TX, Igor RLM, Lee C, Harrison A, Butler R, Dietmann S, Jernej U, Marioni J, Smith CWJ, Göke J, Surani MA. SRSF3 maintains transcriptome integrity in oocytes by regulation of alternative splicing and transposable elements. *Cell Discov*. 2018,4:33.
45. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic integration of splicing within gene regulatory pathways. *Cell*. 2013,152(6):1252-69.
46. Dagueuet, E., Dujardin, G. & Valcarcel, J. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep*. 2015,16, 1640–1655
47. Yohei Hayashi, Masaru Mori, Kaori Igarashi, Keiko Tanaka, Asuka Takehara, Yumi Ito-Matsuoka, Akio Kanai, Nobuo Yaegashi, Tomoyoshi Soga, Yasuhisa Matsui. Proteomic and metabolomic analyses uncover sex-specific regulatory pathways in mouse fetal germline differentiation. *Biol Reprod*. 2020,103(4):717-735.

48. Timothée Revil , Daniel Gaffney, Christel Dias, Jacek Majewski, Loydie A Jerome-Majewska. Alternative splicing is frequent during early embryonic development in mouse. BMC Genomics. 2010,11:399.
49. Yang Q, Wang R, Wei B, Peng C, Wang L, Hu G, Kong D, Du C.Candidate Biomarkers and Molecular Mechanism Investigation for Glioblastoma MultiformeUtilizing WGCNA. Biomed Res Int. 2018,2018:4246703.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.