Supplementary file of “Identification of 4438 novel lincRNAs involved in mouse pre-implantation embryonic development” by *Lv et al.*

Supplementary Methods

The developmental stage specificity measure formulation

We used a probability distribution distance metric related to Jensen-Shannon divergence (JSD) to quantify developmental specificity. The metric quantifies the similarity between expression pattern in a given sample and an extreme pattern that represents that a transcript is expressed in only one stage. The Jensen-Shannon distance $D(a,b)$ between two given normalized expression patterns *a* and *b* is defined as: $D(a,b)=\sqrt{JSD(p_a,p_b)}$, $a=(a_1,\ldots,a_n)$, where p_a and p_b are the abundance distributions of patterns *a* and *b* and $JSD(x,y)$ is the Jensen-Shannon divergence between two probability distributions *x* and *y*, which is defined as: $JSD(x,y)=\frac{1}{2}KLD(x,m)+\frac{1}{2}KLD(y,m)$, where $m=(x+y)/2$ and $KLD(x,y)$ is the Kullback-Leibler divergence between *x* and *y* defined as $KLD(x,y)=\sum_i x_i \log\frac{x_i}{y_i}$. The developmental stage specificity of a transcript’s expression pattern *e*, with respect to stage *t* can then be defined as $DSP(e\mid t)=1-D(e,e^t)$, where e^t is a predefined expression pattern that represents the extreme case in which a transcript is expressed in only one stage. Formally, $e^t=(e_1^t,\ldots,e_n^t), s.t\ \ e_i^t=\begin{cases}1, & if\ i=t\\ 0, & otherwise\end{cases}$. We added a pseudo count of 0.0001 to the abundance distributions to avoid zero. Finally, we define the specificity score of a transcript by the maximal tissue specificity across all predefined expression patterns with expression pattern *e*: $DSP(e)=\max_t DSP(e\mid t)$.

To calculate the specificity of a standardized vector *e* = (0.01, 0.14, 0.85), we need to prepare

three extreme expression patterns: e^1 = (1, 0, 0), e^2 = (0, 1, 0), e^3 = (0, 0, 1). The developmental stage specificity $DSP(e|e^1)$ = 1-$JSD(e, e^1)$ = 0.021, $DSP(e|e^2)$ = 1-$JSD(e, e^2)$ = 0.167 and $DSP(e|e^3)$ = 1-$JSD(e, e^3)$ = 0.723. Thus, $DSP(e)$ = max(0.021,0.167,0.723) = 0.723. For another standardized vector e = (0.3, 0.1, 0.6), the specificity $DSP(e)$ = 0.519.

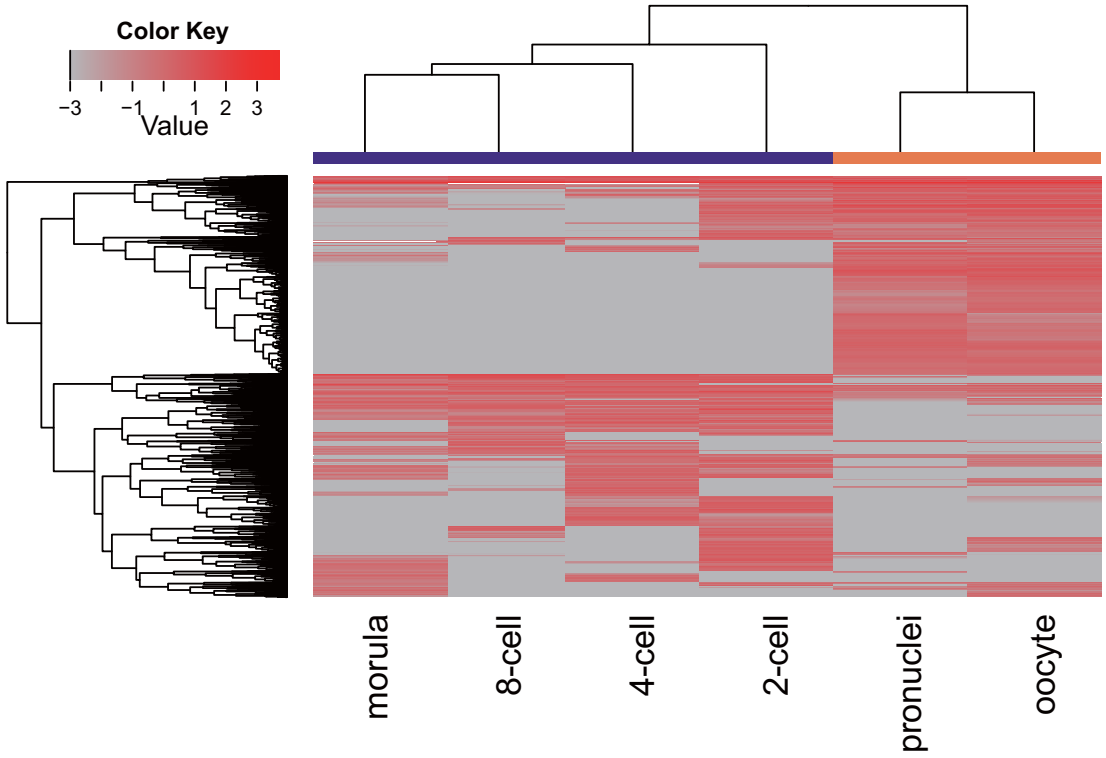

Fig. S1 Expression heatmap of identified novel lincRNAs

Bi-clustering of the log-normalized FPKM estimated by Cufflinks for identified novel lincRNAs across each listed developmental stage

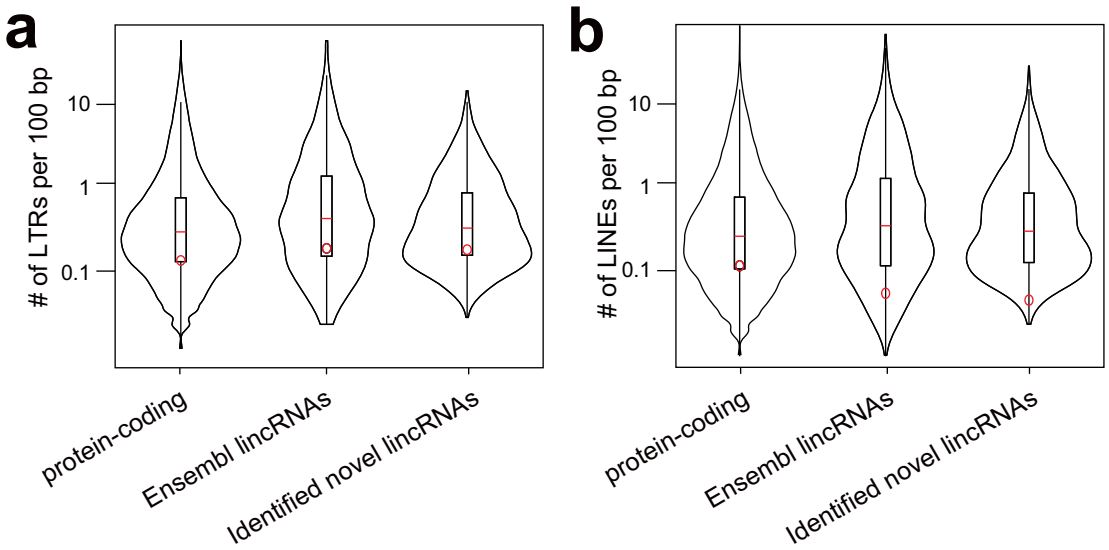

Fig. S2 Characterization of the PED transcriptome by transposons

a Violin plot showing the number of one of the investigated Transposable elements, LTR, per 100bp for lincRNAs and protein-coding genes, and the unfilled circles represent the median values for different gene categories. **b** Violin plot showing the number of one of the investigated Transposable elements, LINE transposon, per 100bp for lincRNAs and protein-coding genes, and the unfilled circles represent the median values for different gene categories

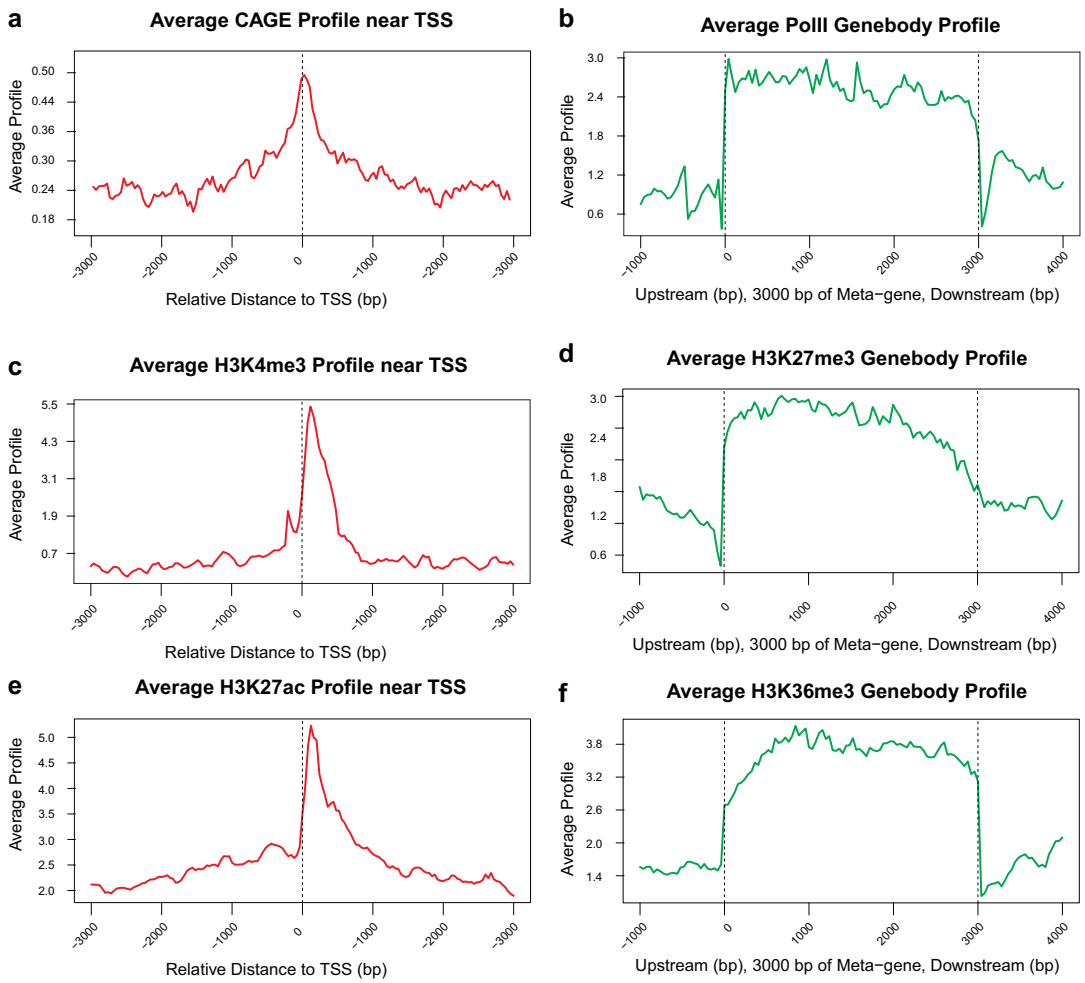

Fig. S3 Novel identified lincRNAs that are expressed in morula are supported by chromatin modifications and CAGE data

a The TSS of novel lincRNAs that are expressed in morula is enriched with CAGE tags over basal levels, where CAGE tag density is aligned around TSS with ±3000 bp extensions. **b** The gene body of novel lincRNAs that are expressed in morula normalized by length of 3000 bp with 1000-bp extension from TSS toward upstream and TTS toward downstream is enriched with PolII tags over basal levels. **c** The TSS of novel lincRNAs is enriched with H3K4me3 signal over basal levels, where H3K4me3 signal density is aligned around TSS with ±3000 bp extensions. **d** The gene body of novel lincRNAs that are expressed in morula normalized by length of 3000 bp with 1000-bp extension from TSS toward upstream and TTS toward downstream is enriched with H3K27me3 signal density over basal levels. **e** The TSS of novel lincRNAs that are expressed in morula is enriched with H3K27ac signal over basal levels, where H3K4me3 signal density is aligned around TSS with ±3000 bp extensions. **f** The gene body of novel lincRNAs that are expressed in morula normalized by length of 3000 bp with 1000-bp extension from TSS toward upstream and TTS toward downstream is enriched with H3K36me3 signal density over basal levels

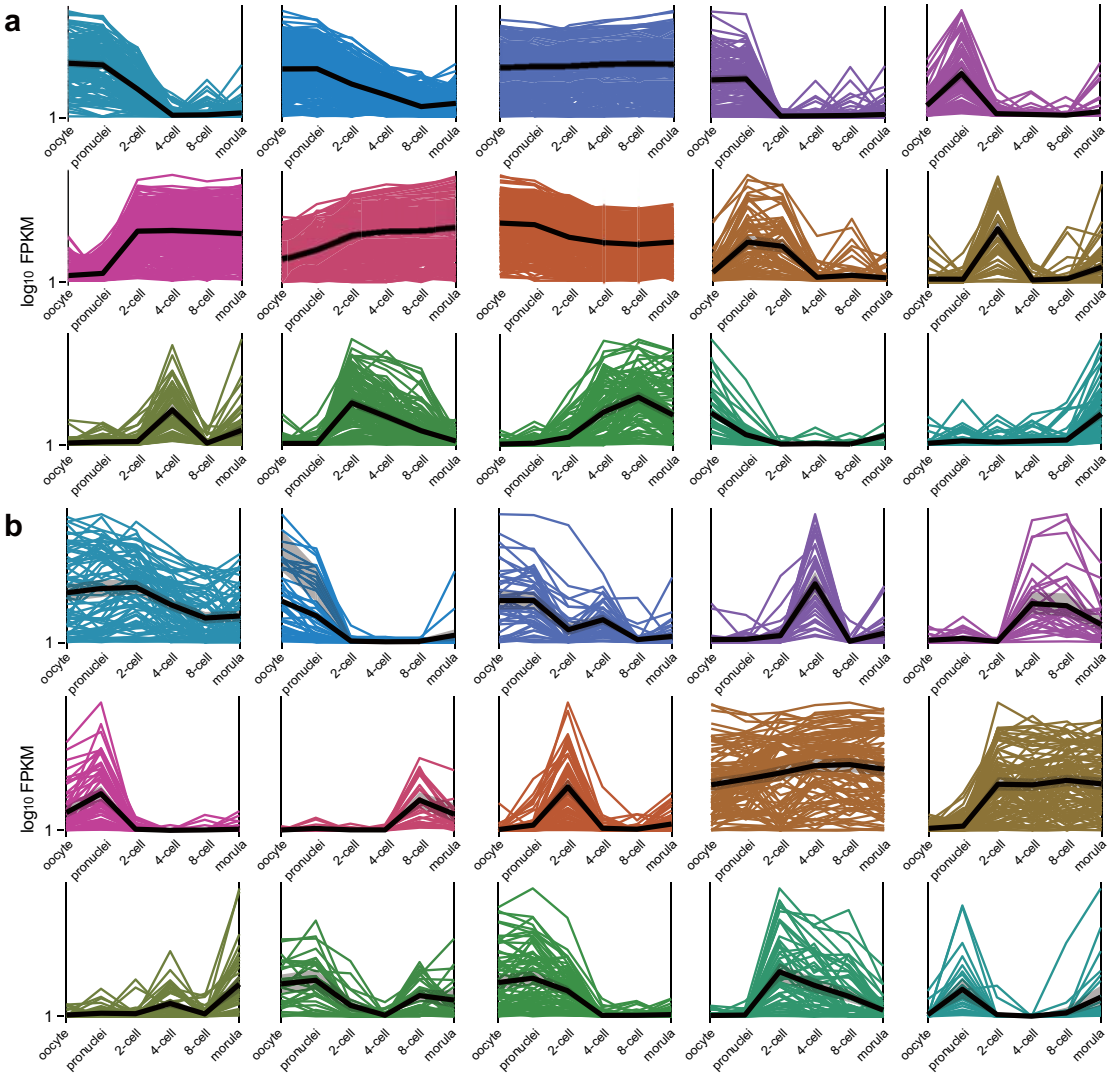

Fig. S4 Expression trends of known genes

Shown are line charts depicting expression patterns of **a** known protein-coding genes and **b** known Ensembl lincRNA genes across different developmental stages. The genes were grouped by the K-means clustering method. X axis represents the six different PED stages; Y axis represents the $\log_{10}$ FPKM expression values of individual genes

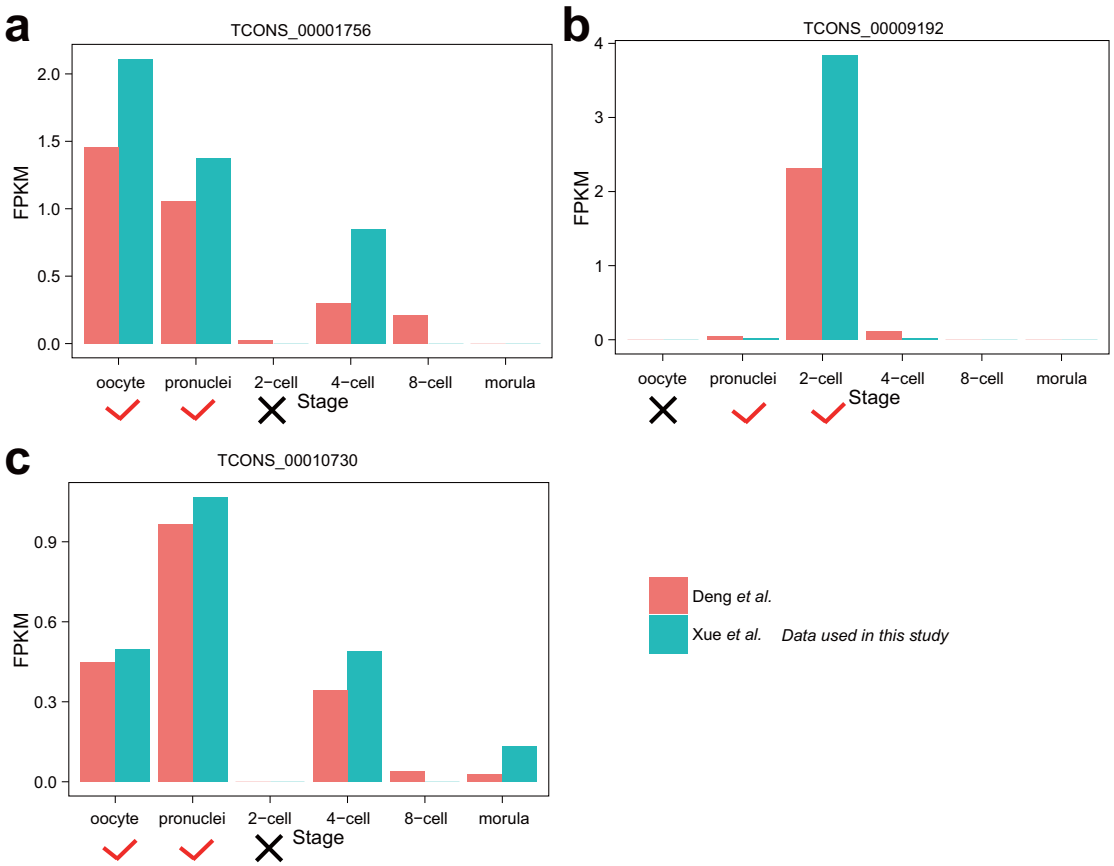

Fig. S5 Validation of three selected lincRNAs that are differentially expressed during zygote first cleavage division

Shown are bar plots of calculated RPKM levels based on two single-cell RNA-seq studies (Xue *et al.* 2013; Deng *et al.* 2014) across different PED stages for **a** *TCONS_00001756*, **b** *TCONS_00009192* and **c** *TCONS_00010730*. X axis represents the six different stages of PED; Y axis represents the FPKM expression levels of specific lincRNAs; red check mark represents that specific lncRNAs can be detected in corresponding stage, while the black cross mark represents that specific lncRNAs cannot be detected in corresponding stage

References

Deng Q, Ramskold D, Reinius B and Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343(6167): 193-196.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu JY, Horvath S and Fan G (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500(7464): 593-597.