Single-cell full-length isoform sequencing reveals an abundance of 3’ prime partial transcripts in maternally inherited RNAs

Chuan-Le Xiao (xiaochuanle@126.com)
Sun Yat-sen University https://orcid.org/0000-0002-4680-0682
Chaoyang Wang
Guangzhou Laboratory
Zhuoxing Shi
Sun Yat-sen University
Qingpei Huang
Guangzhou Laboratory
Kunhua Hu
Proteomics Center, Zhongshan School of Medicine, Sun Yat-sen University,
Rong Liu
Guangzhou Laboratory
Dan Su
Guangzhou Laboratory
Zhuobin Lin
Sun Yat-sen University
Xiaoying Fan
Bioland Laboratory (GuangZhou Regenerative Medicine and Health Guangdong Laboratory)
https://orcid.org/0000-0002-2599-8319

Biological Sciences - Article

Keywords:

Posted Date: June 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1752209/v1

Single-cell full-length isoform sequencing reveals an abundance of 3' prime partial transcripts in maternally inherited RNAs

Chaoyang Wang[1,2,3,6,7], Zhuoxing Shi[1,7], Qingpei Huang[2,7], Kunhua Hu[4,7], Rong Liu[2], Dan Su[2,3,6], Zhuobin Lin[1], Xiaoying Fan[2,3,5,6*], Chuanle Xiao[1,*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong Province, P.R. China

[2]Guangzhou Laboratory, Guangzhou, Guangdong Province, P.R. China

[3]The Bioland Laboratory (GuangZhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, Guangdong Province, P.R. China

[4]Proteomics Center, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong Province, P.R. China

[5]The Fifth Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong Province, P.R. China

[6]The Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Guangdong Province, P.R. China

[7]These authors contributed equally to this work.

*Corresponding authors: Xiaoying Fan (Fan_xiaoying@gzlab.ac.cn), Chuanle Xiao (xiaochuanle@126.com)

Abstract

Isoform expression in preimplantation embryos has been extensively investigated, and many novel isoforms have been identified. However, the regulation patterns of different types of transcripts along the developing stages remains unexplored. We quantified the expression of full-length isoforms in over one hundred single blastomeres from the mouse oocyte to blastocyst stage and found that the 3' prime partial transcripts that lack stop codons were highly accumulated in oocytes and zygotes. A typical 3' prime partial isoform, *Ncl*-S-350, was demonstrated to be not a transcription by-product but to function in the ZGA process. SRSF4 was identified to be responsible for generating these 3' prime partial transcripts in mouse embryonic stem cells (mESCs), and both *Ncl*-S-350 and *Srsf4* overexpression could convert mESCs to a 2-cell (2C)-like state. Our work reveals the

important role of isoform switch regulation in early embryonic development and lays the foundation for an alternative way of acquiring totipotent cells.

Introduction

One gene can be transcribed into multiple kinds of transcripts, namely, isoforms, which are then translated into different proteins. Isoform compositions vary between different cell types and states; therefore, isoform switch is important in determining cell identity[1,2]. In recent years, several third-generation sequencings (TGS) based single-cell RNA-sequencing (scRNA-seq) methods have been developed for direct isoform sequencing[1,3-7]. Among these methods, SCAN-seq has been shown to have high gene detection sensitivity and to enable the discovery of large number of novel transcripts in rare samples, such as preimplantation embryos[4]. Nevertheless, SCAN-seq failed to quantify the absolute abundance of the detected genes and isoforms since it is difficult to specify unique molecular identifiers (UMIs) under the high error rate of Nanopore sequencing[8,9]. Thereby, HIT-scISOseq and MAS-ISO-seq have been used to quantify isoform abundance in single cells with improved data throughput by using the PacBio HiFi sequencing platform[5,7].

Many studies have focused on the molecular regulation of preimplantation embryo development, as it is the basis for reproduction. In particular, the maternal to zygote transition is the foundation of the whole-body development plan[10-13]. Although hundreds of genes have been reported in zygotic genome activation (ZGA), the functional regulators remain largely unclear, not to mention whether the isoform switch participate in the process[14-16].

In this study, we modified the HIT-scISOseq method for low throughput of cells and sequenced the isoforms in single blastomeres of mouse preimplantation embryos. We quantified the isoforms of each gene in each single cell, and different isoform types showed varied proportions among different stages. Specifically, we observed large number of 3' prime partial transcripts (which lack stop codons and generate proteins lacking C-termini) in oocytes and zygotes that quickly degraded from early 2-cell stage (2C) stage. We validated the 3' prime partial nucleolin (*Ncl*) gene transcripts, and surprisingly, the short *Ncl* transcripts may promote ZGA by significantly inducing the expression of *Dux* and other 2C

genes. Moreover, we found that the splicing factor *Srsf4* plays an important role in the expression of 3' prime partial transcripts, which could also induce 2C gene expression in mESCs. Therefore, the 3' prime partial transcripts, which are usually regarded as detrimental by-products, may be essential regulators during preimplantation embryo development, and isoform switch may be a specific regulatory mechanism for the ZGA process.

Results

Characterization of the mouse preimplantation embryos by gene and isoform expression

To detect gene isoforms in each blastomere of the mouse metaphase II oocytes and preimplantation embryos, we amplified RNAs in each single cell with a 10x gel bead using the Smart-seq2 protocol[5,17]. Then, the amplified cDNAs of different cells were pooled for further ligation and PacBio library construction following the HIT-scISOseq method[5]. Meanwhile, the corresponding barcode sequence of each cell was pre-determined through Sanger sequencing of the cDNAs (Extended Data Fig. 1a). Expression data for a total of 161 single-cell isoforms were obtained from three batches, comprising the mouse oocyte, zygote, early 2-cell (E2C), late 2-cell (L2C), 4-cell, 8-cell, 16-cell, 32-cell, and blastocyst stages and mESCs (Extended Data Table 1). Each sequencing batch generated about 5 million circular consensus sequencing (CCS) reads, with the average length around 4kb, indicating ligation of 2-3 cDNA molecules in most cases. After data splitting and mapping, about 90% of the full-length isoforms could be correctly assigned to cells (Extended Data Table 1). In this way, the samples from each stage could be sequenced with a relatively sufficient depth (Extended Data Fig. 1b). To evaluate the accuracy of measuring absolute numbers of isoforms with our procedure, we also amplified ERCC and SIRV spike-ins[18]. At the gene level, we observed high correlation values between the added molecules and the detected UMI counts (Extended Data Fig. 1c). At the isoform level, different isoforms of the same gene could be correctly specified without any false match (Extended Data Fig. 1d). These results indicate that our workflow accurately measures the abundances of the transcripts in each single cell.

As expected, the mouse oocyte and zygote contained many more RNA molecules than later stage blastomeres as a result of subsequent maternally inherited RNA degradation. The number of transcript molecules was well correlated (R=0.96) with the number of genes expressed in the cells (Fig. 1a). We used gene expression data and isoform expression data to perform principal component analysis (PCA) of all the cells. Both types of data generated similar PCA results and could be used to clearly separate blastomeres of different stages (Fig. 1b, c). In both sets of PCA results, the oocyte and zygote showed similar expression patterns, the L2C and 4-cell stages could be grouped together, the 8-cell, 16-cell and 32-cell stages were close, and the blastocyst cells were analogous to the mESCs. We further extracted stage-specific genes and transcripts with the same criterion and obtained 3867 genes and 6819 isoforms, respectively. Each list could be divided into six corresponding clusters based on their expression patterns across all embryonic stages (Fig. 1d, e). To be specific, Cluster 1 (C1) transcripts were highly abundant in oocytes and zygotes, subsequently degraded from E2C stage. Cluster 2 (C2) included transcripts only upregulated in E2C stage. Cluster 3 (C3), cluster 4 (C4) and cluster 5 (C5) transcripts were highly expressed in the L2C to 4-cell stages, 8-cell to 32-cell stages and blastocyst stages, respectively. The mESC-specific transcripts were in cluster 6 (C6). In addition, we compared the genes and isoforms in each pair of matched clusters. More number of RNAs could be identified in each cluster at the isoform level, and most of the isoforms showed consistency with the genes (Fig. 1f, Extended Data Table 2). Thus, the single-cell isoform expression data could be directly used to illustrate cellular heterogeneity and distinguish different types of cells as single-cell gene expression does.

To investigate the relationship between gene and isoform expression during mouse preimplantation development, we divided genes into six groups based on the number of isoform types they expressed (Extended Data Fig. 2a). Although majority of genes expressed only one type of isoform across different stages, more genes expressed multiple types of isoforms in the earlier stages. About 60% of genes in mouse oocytes and zygotes expressed more than one type of isoforms, and nearly 20% of genes were detected with more than five types of isoforms. In comparison, approximately 70% of genes in mESCs expressed only one type of isoform, and few genes expressed over five types of isoforms

(Extended Data Fig. 2a). The same isoform expression characteristics could also be observed with SCAN-seq data (Extended Data Fig. 2b), suggesting rich isoform diversity in early mouse embryos. In addition, the genes expressing more types of isoforms were detected with higher expression levels in both our data and SCAN-seq data (Extended Data Fig. 2c, d). To assess the isoform dominant level in each gene expressing multiple types of isoforms, we calculated the ratio of the UMI number of the major isoform to the total UMI number of the corresponding gene. The major isoform ratios increased from early to late embryonic stages, especially after the ZGA process (Extended Data Fig. 2e). In comparison, the major isoforms accounted for 90% of most genes in mESCs, indicating a dominant isoform expression pattern and less isoform diversity in these cells.

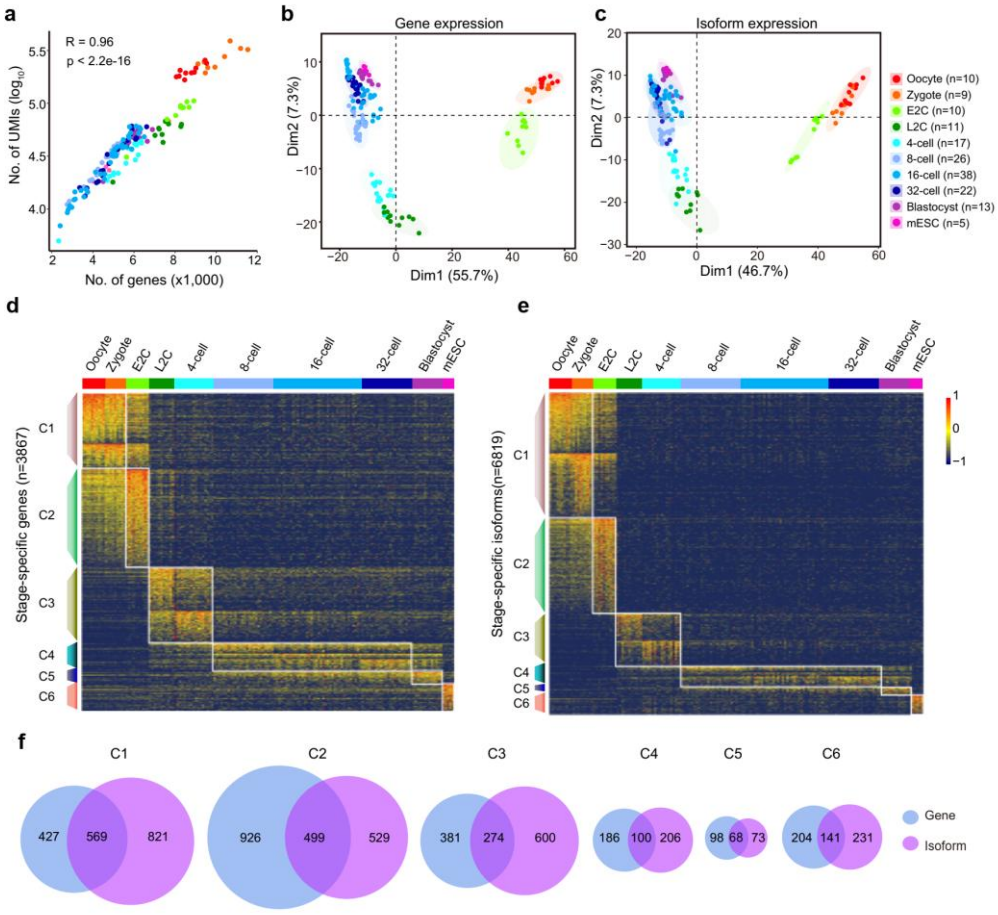


Fig. 1| Gene and isoform expression in the mouse preimplantation embryos

a, Number of genes and UMIs in each cell at different stages. **b**-**c**, PCA plot of all the blastomeres and mESCs based on gene expression (b) and isoform expression (c). **d-e**, Heatmap of stage-specific genes (d) and isoforms (e). **f**, Venn plot of pairwise groups of stage-specific genes and isoform-corresponding genes.

Large abundance of 3’ prime partial transcripts was observed in mouse oocytes and zygotes

According to the integrity of putative corresponding open reading frames (ORFs), the transcripts were divided into 5 types: complete, representing isoforms coding the complete ORFs of the reference genes; 3’ prime partial transcripts and 5’ prime partial transcripts, missing the stop codon section and start codon section, respectively; internal, predicted with proteins lacking both ends; and others, including isoforms located outside the reference ORFs[19] (Fig. 2a). As expected, the complete transcripts had the longest lengths, and the internal transcripts were the shortest (Extended Data Fig. 3a). Nevertheless, the predicted protein length was similar for the three incomplete transcript types (Extended Data Fig. 3b). Intriguingly, we found that the 3’ prime partial transcripts were highly expressed in oocytes and zygotes and then dramatically decreased from the E2C stage (Fig. 2b, c). The same expression pattern was also observed in the SCAN-seq data (Extended Data Fig. 3c, d). Then, we performed Gene Ontology (GO) analysis on genes detected with 3’ prime partial transcripts (Extended Data Table 3). These genes were enriched in the pathways of RNA processing, cell cycle checkpoint, ribonucleoprotein complex biogenesis, DNA metabolic process, chromatin organization, etc. (Extended Data Fig. 3e), all of which have been reported to play essential roles in mouse and human preimplantation embryo development[10,13,20-23]. In addition, we found that 82% of these 3’ prime partial transcripts ended within the exon, while only 44% of the 5’ prime partial transcripts ended within the exon (Extended Data Fig. 3f).

We further chose some candidates to validate the enrichment of 3’ prime partial transcripts in the maternally inherited RNAs. For example, the *Ncl* gene was detected with 6 isoform types that could be assigned to three categories based on how many RNA recognition motifs (RRM) they contained (Fig. 2d, left). All 6 isoform types were observed in the reverse transcription and PCR (RT–PCR) products of mouse oocytes (Fig. 2d, right). Meanwhile, the short isoforms appeared to be much more abundant than the full-length isoform (*Ncl*-FL-71) in mouse oocytes. Then, we calculated the abundance of each category of *Ncl* isoform along developing stages. *Ncl* was first downregulated in the E2C

stage and then increased from the L2C stage at the gene level (Fig. 2e). According to changes in the abundances of different isoforms, we found that the two categories of short *Ncl* isoforms were highly expressed in oocytes and zygotes and then almost disappeared. In comparison, the full-length *Ncl* isoform only showed low expression in the maternal RNAs and was largely upregulated after ZGA (Fig. 2e). This finding highlights the isoform switch of the *Ncl* gene during the ZGA process; and such regulation approaches are masked in gene-level analysis. To demonstrate this, we performed reverse transcription and real-time quantitative PCR (RT–qPCR) using primers targeting all the *Ncl* isoform types and only the full-length type, respectively, and calculated the relative percentages of the full-length isoform in different stages of mouse preimplantation embryos. As expected, less than 20% of the *Ncl* transcripts were full length in oocytes when we set the full-length relative ratio as 100% at the morula stage (Fig. 2f). The detected 3' prime partial transcripts of some other genes related to RNA processing (*Sf3b2, Srpk1*) and protein translation and transporting (*Dnajc3, Hsp90aa1*) were also revealed by gel analysis of mouse oocyte RT–PCR products (Extended data Fig. 4a).

The 3' prime partial transcripts lack the stop codon by definition. We further sequenced the amplified 3' prime partial transcripts of *Ncl* and genes in Extended data Fig. 4a by Sanger sequencing for validation. Each predicted isoform was confirmed to have a poly(A) tail without a stop codon (Extended data Fig. 4b). Usually, transcripts without stop codons are quickly degraded by a nonstop decay (NSD) mechanism in eukaryotic cells[24,25]. We analysed the expression of *Pelota, Hbs1* and *Abce1*, the three indispensable regulators triggering the NSD process (Extended data Fig. 4c). Both *Pelota* and *Abce1* were expressed at low levels in oocytes and zygotes, indicating the NSD pathway shall be in quite low activity before ZGA process. In this way, the 3' prime partial transcripts can be largely accumulated in the maternal content.

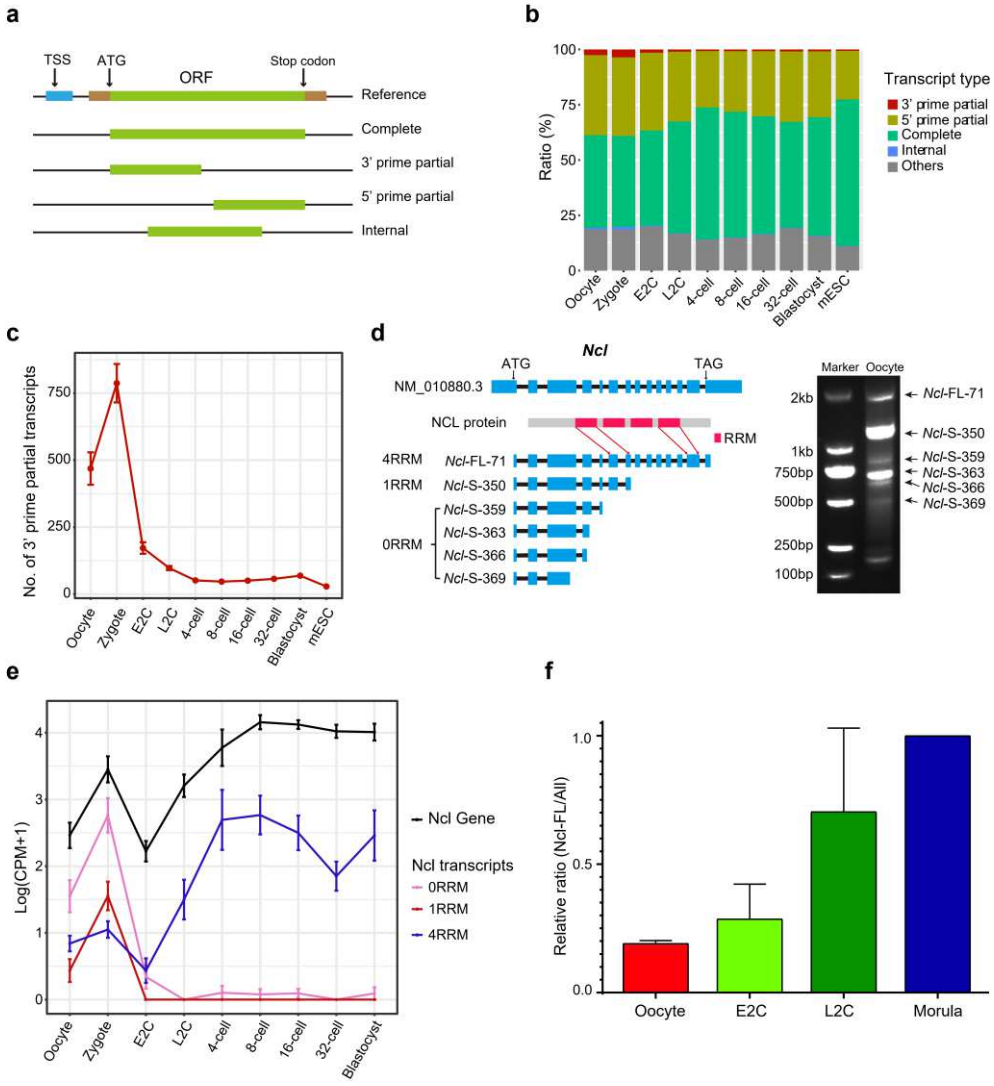


Fig. 2| Expression patterns of different types of transcripts during mouse preimplantation embryonic development

a, Schematic diagram of different transcript types. **b**, Ratios of each type of transcripts at different stages. **c**, Number of 3' prime partial transcripts detected at each stage. **d**, Schematic diagram of *Ncl* isoforms (left) and gel picture showing these isoforms by RT-PCR of *Ncl* in mouse oocytes (right). **e**, Expression levels of each category of *Ncl* isoforms and the *Ncl* gene at different stages. **f**, The relative ratios of *Ncl* full-length isoforms at different stages detected by RT-qPCR. The relative ratio of NCL-FL/All in morulae was set as 1.0.

The 3’ prime partial *Ncl* transcript (*Ncl*-S-350) induces the expression of 2C genes

NCL *encoded* by the *Ncl* gene has been demonstrated to participate in a wide range of cellular programs, including ribosome biogenesis, chromatin organization and stability, and DNA and RNA metabolism[26]. This gene is highly expressed in mESCs, and it has been proven that NCL forms a complex with KAP1/TRIM28 and *LINE1* RNA to repress the expression of *Dux*, the master activator of the 2C transcription program[27-31]. Knocking down *Ncl* in mESCs by siRNA had been reported to lead to activation of the 2C transcription program characterized by the expression of a series of 2C genes, such as *Dux*, *MERVL*, *Zscan4C* and *Tsctv1*[27,32]. There were very few 3’ prime partial transcripts in mESCs, including *Ncl*-S-350, which was highly detected in the maternal RNA content (Fig. 2b, 2c). To determine whether these 3’ prime partial transcripts are functional, we overexpressed *Ncl*-S-350 in mESCs (Fig. 3a). Surprisingly, *Ncl*-S-350 overexpressing cells also significantly upregulated 2C genes, including *Dux*, *MERVL*, and *Zscan4c* (Fig. 3b). This finding was consistent with the results when *Ncl* was knocked down in mESCs (Fig. 3c)[27,32].

The direct interaction between *LINE1* RNA and NCL has been confirmed *in vitro* and *in vivo*[33,34]. Moreover, the full integrity of the first and second RRM domains is required to bind RNA[35]. As NCL inhibits *Dux* expression under the guidance of *LINE1* RNA[27], we deduced that the NCL-S-350 lacks the ability to bind *LINE1* RNA compared to the full-length isoform (NCL-FL). Therefore, the NCL-S-350 protein may function as an antagonist of the NCL-FL-71 protein to form a complex with KAP1/TRIM28 without *LINE1* RNA. To verify this assumption, we did co-immunoprecipitation (Co-IP) assay and the result confirmed that NCL-S-350 protein can bind KAP1 as the NCL-FL proteins do (Fig. 3d). The results imply that the 3’ prime partial transcripts shall not be transcription by-products and may play an important role in preimplantation embryo development.

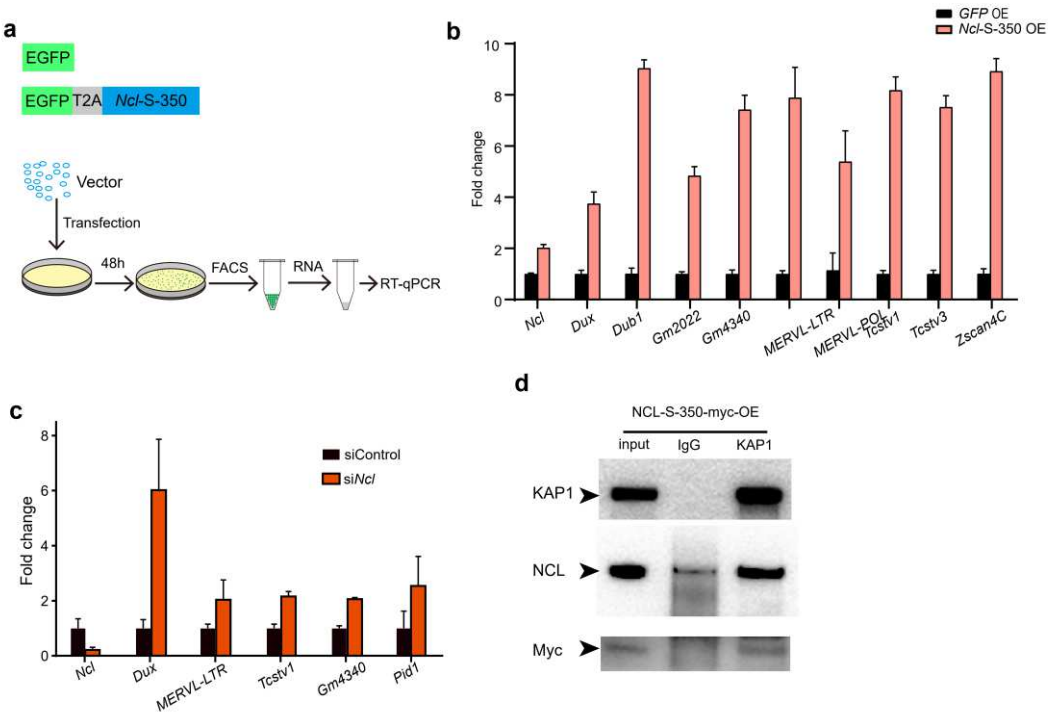


Fig. 3| Overexpression of the 3' prime partial transcript *Ncl*-S-350 induced 2C gene expression in mESCs

a, Overview of the overexpression workflow in mESCs. **b-c**, Bloxplot showing increased expression levels of *Dux* and other 2C genes by RT-qPCR after overexpressing *Ncl*-S-350 (b) or knocking down *Ncl* using siRNA (c) in mESCs (n=3). **d**, Co-IP assay showing that NCL-S-350 can combine with KAP1 as NCL-FL-71 does.

Srsf4 promotes the accumulation of 3' prime partial transcripts

The 3' prime partial transcripts in the maternally inherited RNAs mostly ended within the exon region (Extended Data Fig. 3f) and were likely not regulated by spliceosomes, which function in introns. Exon splicing requires serine/arginine-rich splicing factors (SRs)[36,37]. Therefore, we speculated that the generation of 3' prime partial transcripts rely on SRs. The SR family consists of 12 members in mammals[38]. We checked the expression patterns of each SR gene in the mouse preimplantation embryo samples and noticed that *Srsf4* was highly expressed in oocytes and zygotes and then downregulated at the later stages (Fig. 4a, Extended data Fig. 5). The same expression patterns of the SR genes were also observed in previous NGS data (Fig. 4a, Extended data Fig. 6)[39]. We further analysed the motifs using the 100 bp sequence flanking the end point of all the 3' prime

247 partial transcripts (Fig. 4b). Interestingly, AGAAAA is consistent with the conserved binding
248 motif of SRSF4, which has been reported to work on *Ncl* mRNA[40]. We checked the
249 distribution of the AGAAAA motif on all the exons of *Ncl*, and its locations matched perfectly
250 to the end sites of those 3' prime partial transcripts of *Ncl* (Extended data Fig. 7a).

251 In order to check the function of SRSF4 in producing the 3' prime partial transcripts,
252 we overexpressed *Srsf4* in mESCs. Subsequently, we amplified the *Ncl* transcripts and
253 examined the isoform types on the gel. Intriguingly, compared to the phenomenon that *Ncl*-
254 FL-71 was dominant in normal mESCs, *Ncl*-S-350 became the most abundant isoform type
255 in *Srsf4*-overexpressing mESCs (Fig. 4c). The other 3' prime partial transcripts of *Ncl* were
256 also increased. Moreover, the expression pattern of *Ncl* isoforms in *Srsf4*-overexpressing
257 mESCs was almost identical to that in mouse oocytes (Fig. 4c). These results suggested
258 that *Srsf4* regulates the generation of the 3' prime partial transcripts of *Ncl.* We further
259 detected the expression of 2C genes in *Srsf4*-overexpressing mESCs by RT-qPCR. As
260 expected, *Dux, MERVL, Zscan4c* and other 2C genes were all significantly upregulated,
261 which was consistent with the results in mESCs overexpressing *Ncl*-S-350 (Fig. 4d).

262 To further investigate the targets of SRSF4 and whether the cells acquired totipotency
263 when overexpressed with the 3' prime partial transcripts, we performed single-cell full-
264 length isoform sequencing on *Srsf4*-overexpressing mESCs (*Srsf4* OE), *Ncl*-S-350-
265 overexpressing mESCs (*Ncl*-S-350 OE) and GFP-overexpressing control mESCs (*GFP*
266 OE). The *Srsf4* OE and *Ncl*-S-350 OE cells were almost the same with each other but
267 different from the *GFP* OE control mESCs (Extended Data Fig. 7b, c). Overexpression of
268 both *Srsf4* and *Ncl*-S-350 led to elevated expression of totipotent genes in mESCs (Fig.
269 4e, Extended Data Table 4). In addition, most minor and major ZGA genes were also
270 upregulated (Extended Data Fig. 7d, Extended Data Table 4). To check the states of the
271 two groups of cells, we clustered them with the embryo samples and previously reported
272 totipotent-like stem cells (TLSCs)[16]. Interestingly, both *Srsf4* OE and *Ncl*-S-350 OE cells
273 were close to the TLSCs, all showing similarity to the 2C embryo (Fig. 4f).

274 Notably, the proportion of 3' prime partial transcripts increased from 7% in the *GFP*
275 OE group to about 19% in the *Srsf4* OE and *Ncl*-S-350 OE groups (Fig. 4g). Genes
276 generating 3' prime partial transcripts under SRSF4 were related to RNA processing,

mitotic cell cycle, chromatin organization, etc., which was similar to those enriched in the mouse preimplantation embryos (Extended Data Table 5 and Fig. 7e). Moreover, the isoform diversity increased in the *Srsf4* OE and *Ncl*-S-350 OE groups, as 43.7% and 46.2% of genes in the two groups expressed more than 5 isoform types, respectively, which was much higher than the ratio (33.9%) in the *GFP* OE group (Fig. 4h). This finding also resembled the large isoform diversity in oocytes and early embryos (Extended Data Fig. 2a, b). Therefore, *Srsf4* may play an essential role in the generation of 3' prime partial transcripts, and both *Srsf4* and *Ncl*-S-350 may participate in cellular identity determination in mouse oocytes and early embryos.

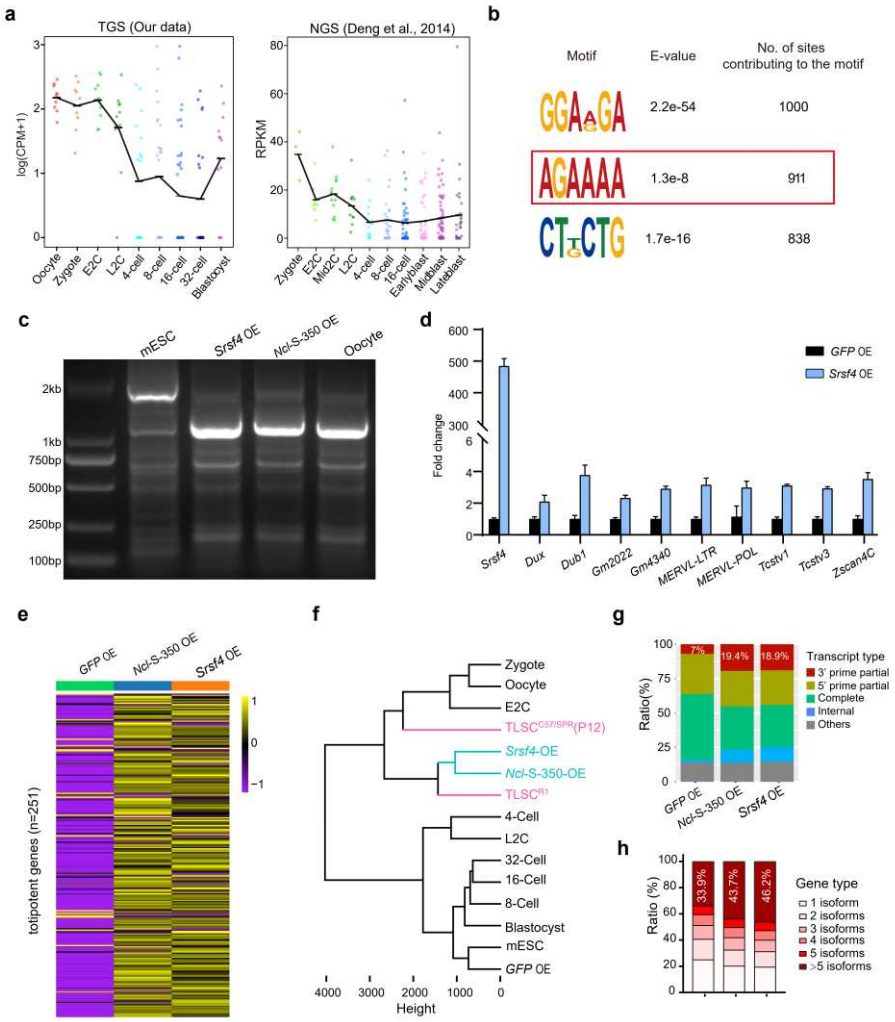


Fig. 4| *Srsf4* regulates the production of the 3'prime partial transcripts of *Ncl* and other genes

a, Expression levels of *Srsf4* at different embryonic stages in our data and in the NGS data by Deng *et al.*, 2014[39]. **b**, The significant motifs in the 100 bp sequences flanking the end sites of the 3' prime partial transcripts. The red box highlights the one consistent with the SRSF4 binding motif. **c**, Gel analysis of RT-PCR products of normal mESCs, *Srsf4* OE mESCs, *Ncl*-S-350 OE mESCs and mouse MII oocytes. **d**, Bloxplot showing increased expression levels of *Ncl*, *Dux* and other 2C genes by RT-qPCR after overexpressing *Srsf4* in mESCs. **e**, Heatmap showing generally elevated expression of totipotent genes in *Srsf4* OE and *Ncl*-S-350 OE cells compared with control *GFP* OE cells. **f**, Dendrogram displaying the *Srsf4* OE and *Ncl*-S-350 OE cells were similar to previously reported TLSCs and 2-cell embryos in gene expression. **g**, Ratio of each type of transcript in the three groups of cells. **h**, The ratios of genes containing different numbers of isoform types in the three groups of cells.

Discussion

An abundance of novel transcripts and splicing events in preimplantation embryos have been annotated by TGS-based full-length isoform sequencing at either the bulk or single-cell level[4,14,41]. However, it remains unclear how different types of isoforms regulate the developmental process. In the present study, by dividing the transcripts into subtypes according to their coding characteristics, we found large number of 3' prime partial transcripts, which lack stop codons, in mouse MII oocytes and zygotes (Fig. 2b). Defects in the NSD pathway in the maternal contents enable these transcripts to be maintained. This type of transcript has been largely studied in cancers and is considered an oncogenic factor[42]. While in early embryos, these transcripts might be important for promoting the ZGA process, as their host genes were highly enriched in RNA processing, cell cycle checkpoint, ribonucleoprotein complex biogenesis, DNA metabolic process and chromatin organization pathways (Extended Data Fig. 3e).

Taking *Ncl* as an example, we proved that the 3' prime partial transcripts of *Ncl* promoted the expression of *Dux* and other 2C genes, possibly by competing with the full-length isoform to form the NCL/Kap1 complex[27]. However, they lost the capacity to bind *LINE1* RNA to locate the repressive complex to the *Dux* promoter (Fig. 5). After ZGA, these transcripts are degraded rapidly, and the full-length isoform is upregulated to suppress *Dux* again. Such isoform switch regulation is blinded through single-cell gene expression analysis, emphasizing the importance of performing isoform sequencing.

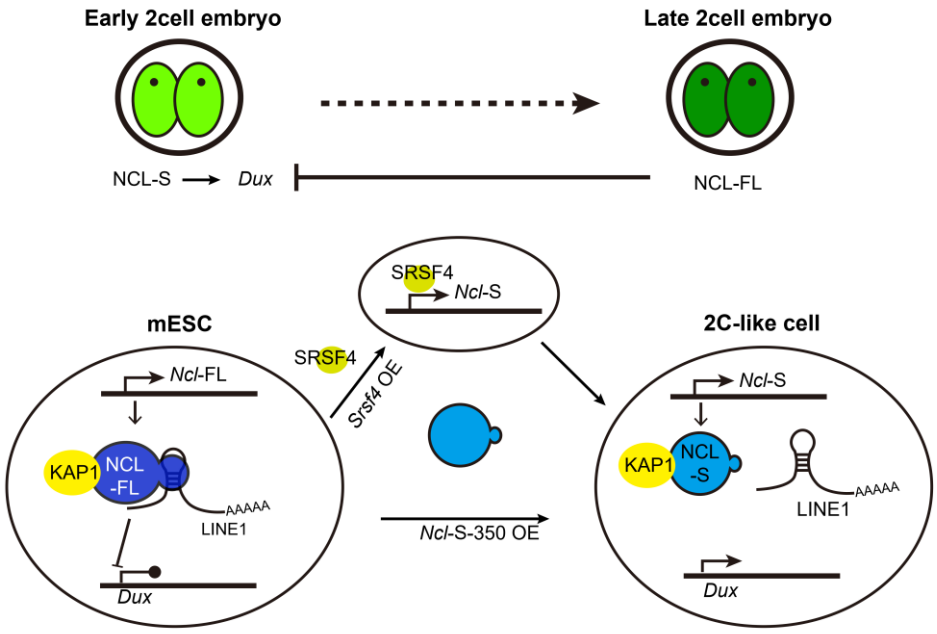


Fig. 5| The hypothesis of the regulatory mechanism of *Dux* by *Ncl*-S-350

In early 2-cell stage, the numerous existing 3' prime partial types of NCL could enhance the activation of *Dux*, while in late 2-cell stage, these transcripts degrade rapidly, and the increased full-length NCL suppresses *Dux* expression. In mESCs, *Dux* is suppressed by full-length NCL[27]. Overexpression of 3' prime partial type *Ncl*-S-350 directly or by *Srsf4* overexpression results in competition with the full-length NCL form complex with KAP1 but lacks LINE1 as the guidance to suppress *Dux*.

A recent study demonstrated that spliceosomal repression induces totipotency in mESCs[43], indicating that alternative splicing is essential for modulating cell fate. However, the PlaB they used to inhibit SF3B, which forms spliceosomes to splice the intron region[43-45], and the 3' prime partial transcripts in the maternal RNAs primarily ended within the exon region. We proved that SRSF4 can produce the 3' prime partial transcripts of both *Ncl* and many other genes, inducing a 2C-like state in mESCs (Fig. 4). We also noticed that the overexpression of *Ncl*-S-350 enhanced the expression of *Srsf4* (data not shown). This could explain why the *Srsf4* OE and *Ncl*-S-350 OE cells were almost the same as each other and similar to TLSCs from a previous study. This may promote a new way for the *in vitro* culture of totipotent cells, and such an assumption needs further experimental exploration and verification.

We extrapolated that the large number of 3' prime partial transcripts in maternally inherited RNAs are not transcription by-products. These transcripts may play functional roles in the preimplantation embryo development. In current work, we investigated the function of *Ncl*, while there are many other genes worthy of further study. Our work provides new insights into the mechanisms regulating early embryonic development.

References

1 Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11**, 4025, doi:10.1038/s41467-020-17800-6 (2020).

2 Joglekar, A. *et al.* A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**, 463, doi:10.1038/s41467-020-20343-5 (2021).

3 Lebrigand, K. *et al.* The spatial landscape of gene expression isoforms in tissue sections. *bioRxiv*, 2020.2008.2024.252296, doi:10.1101/2020.08.24.252296 (2022).

4 Fan, X. *et al.* Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLoS Biol* **18**, e3001017, doi:10.1371/journal.pbio.3001017 (2020).

5 Zheng, Y.-F. *et al.* HIT-scISOseq: High-throughput and High-accuracy Single-cell Full-length Isoform Sequencing for Corneal Epithelium. *bioRxiv*, 2020.2007.2027.222349, doi:10.1101/2020.07.27.222349 (2020).

6 Oguchi, Y., Ozaki, Y., Abdelmoez, M. N. & Shintaku, H. NanoSINC-seq dissects the isoform diversity in subcellular compartments of single cells. *Science Advances* **7**, eabe0317, doi:doi:10.1126/sciadv.abe0317 (2021).

7 Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv*, 2021.2010.2001.462818, doi:10.1101/2021.10.01.462818 (2021).

8 Liu, Z. *et al.* Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol* **23**, 68, doi:10.1186/s13059-022-02636-8 (2022).

9 Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics* **14**, 265-279, doi:10.1016/j.gpb.2016.05.004 (2016).

10 Jukam, D., Shariati, S. A. M. & Skotheim, J. M. Zygotic Genome Activation in Vertebrates. *Dev Cell* **42**, 316-332, doi:10.1016/j.devcel.2017.07.026 (2017).

11 Tadros, W. & Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033-3042, doi:10.1242/dev.033183 (2009).

12 Vastenhouw, N. L., Cao, W. X. & Lipshitz, H. D. The maternal-to-zygotic transition revisited. *Development* **146**, doi:10.1242/dev.161471 (2019).

13 Zhao, L. W. *et al.* Nuclear poly(A) binding protein 1 (PABPN1) mediates zygotic genome activation-dependent maternal mRNA clearance during mouse early embryonic development. *Nucleic Acids Res* **50**, 458-472, doi:10.1093/nar/gkab1213 (2022).

14 Qiao, Y. *et al.* High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nat Commun* **11**, 2653, doi:10.1038/s41467-020-16444-w (2020).

15 Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57-63, doi:10.1038/nature11244 (2012).

16 Yang, M. *et al.* Chemical-induced chromatin remodeling reprograms mouse ESCs to totipotent-like stem cells. *Cell Stem Cell* **29**, 400-418 e413, doi:10.1016/j.stem.2022.01.010 (2022).

17 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells.

Nat Methods **10**, 1096-1098, doi:10.1038/nmeth.2639 (2013).
18 Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods* **13**, 792-798, doi:10.1038/nmeth.3958 (2016).
19 Saha, S., Matthews, D. A. & Bessant, C. High throughput discovery of protein variants using proteomics informed by transcriptomics. *Nucleic Acids Res* **46**, 4893-4902, doi:10.1093/nar/gky295 (2018).
20 Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593-597, doi:10.1038/nature12364 (2013).
21 Yang, Y. *et al.* RNA 5-Methylcytosine Facilitates the Maternal-to-Zygotic Transition by Preventing Maternal mRNA Decay. *Mol Cell* **75**, 1188-1202 e1111, doi:10.1016/j.molcel.2019.06.033 (2019).
22 Sha, Q. Q. *et al.* Characterization of zygotic genome activation-dependent maternal mRNA clearance in mouse. *Nucleic Acids Res* **48**, 879-894, doi:10.1093/nar/gkz1111 (2020).
23 Fu, X., Zhang, C. & Zhang, Y. Epigenetic regulation of mouse preimplantation embryo development. *Curr Opin Genet Dev* **64**, 13-20, doi:10.1016/j.gde.2020.05.015 (2020).
24 Powers, K. T., Szeto, J. A. & Schaffitzel, C. New insights into no-go, non-stop and nonsense-mediated mRNA decay complexes. *Curr Opin Struct Biol* **65**, 110-118, doi:10.1016/j.sbi.2020.06.011 (2020).
25 Morris, C., Cluet, D. & Ricci, E. P. Ribosome dynamics and mRNA turnover, a complex relationship under constant cellular scrutiny. *Wiley Interdiscip Rev RNA* **12**, e1658, doi:10.1002/wrna.1658 (2021).
26 Jia, W., Yao, Z., Zhao, J., Guan, Q. & Gao, L. New perspectives of physiological and pathological functions of nucleolin (NCL). *Life Sciences* **186**, 1-10, doi:10.1016/j.lfs.2017.07.025 (2017).
27 Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405 e319, doi:10.1016/j.cell.2018.05.043 (2018).
28 Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat Genet* **49**, 925-934, doi:10.1038/ng.3844 (2017).
29 Whiddon, J. L., Langford, A. T., Wong, C. J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nat Genet* **49**, 935-940, doi:10.1038/ng.3846 (2017).
30 De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**, 941-945, doi:10.1038/ng.3858 (2017).
31 Ren, W. *et al.* DUX: One Transcription Factor Controls 2-Cell-like Fate. *Int J Mol Sci* **23**, doi:10.3390/ijms23042067 (2022).
32 Sun, Z. *et al.* LIN28 coordinately promotes nucleolar/ribosomal functions and represses the 2C-like transcriptional program in pluripotent stem cells. *Protein Cell*, doi:10.1007/s13238-021-00864-5 (2021).
33 Peddigari, S., Li, P. W., Rabe, J. L. & Martin, S. L. hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. *Nucleic Acids Res* **41**, 575-585, doi:10.1093/nar/gks1075 (2013).
34 Moldovan, J. B. & Moran, J. V. The Zinc-Finger Antiviral Protein ZAP Inhibits LINE and

Alu Retrotransposition. *PLoS Genet* **11**, e1005121, doi:10.1371/journal.pgen.1005121 (2015).

35 Serin, G. *et al.* Two RNA-binding domains determine the RNA-binding specificity of nucleolin. *J Biol Chem* **272**, 13109-13116, doi:10.1074/jbc.272.20.13109 (1997).

36 Zheng, X. *et al.* Serine/arginine-rich splicing factors: the bridge linking alternative splicing and cancer. *Int J Biol Sci* **16**, 2442-2453, doi:10.7150/ijbs.46751 (2020).

37 Jang, H. N. *et al.* Binding of SRSF4 to a novel enhancer modulates splicing of exon 6 of Fas pre-mRNA. *Biochem Biophys Res Commun* **506**, 703-708, doi:10.1016/j.bbrc.2018.10.123 (2018).

38 Long, J. C. & Caceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**, 15-27, doi:10.1042/BJ20081501 (2009).

39 Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193-196, doi:10.1126/science.1245316 (2014).

40 Anko, M. L. *et al.* The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* **13**, R17, doi:10.1186/gb-2012-13-3-r17 (2012).

41 Berrens, R. V. *et al.* Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nat Biotechnol*, doi:10.1038/s41587-021-01093-1 (2021).

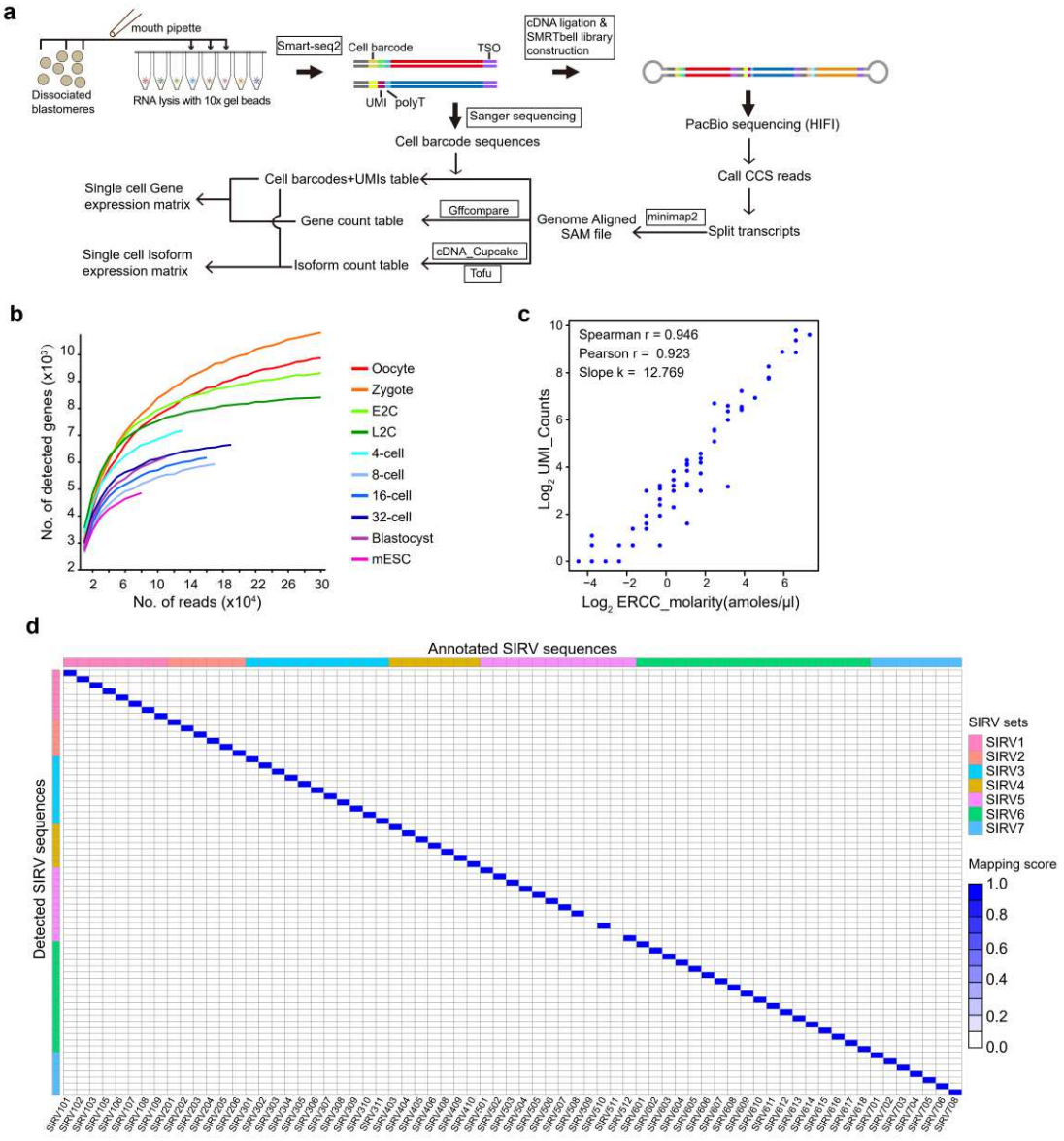
42 Mohanan, N. K., Shaji, F., Koshre, G. R. & Laishram, R. S. Alternative polyadenylation: An enigma of transcript length variation in health and disease. *WIREs RNA* **13**, doi:10.1002/wrna.1692 (2021).

43 Shen, H. *et al.* Mouse totipotent stem cells captured and maintained through spliceosomal repression. *Cell* **184**, 2843-2859 e2820, doi:10.1016/j.cell.2021.04.020 (2021).

44 Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nat Chem Biol* **3**, 570-575, doi:10.1038/nchembio.2007.16 (2007).

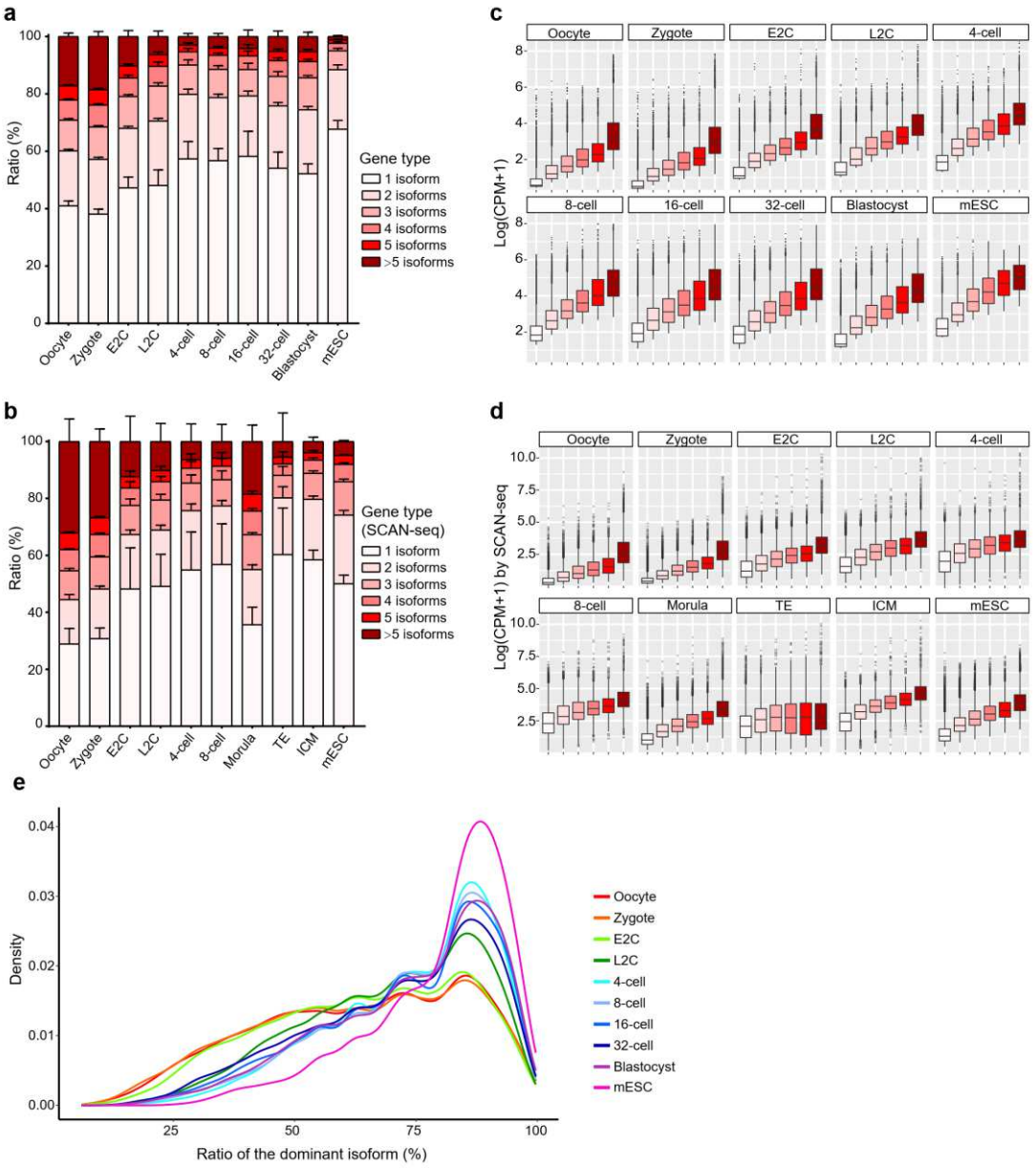
45 Sun, C. The SF3b complex: splicing and beyond. *Cell Mol Life Sci* **77**, 3583-3595, doi:10.1007/s00018-020-03493-z (2020).

Extended Data Fig



Extended Data Fig. 1| Quality evaluation of the single-cell isoform expression data

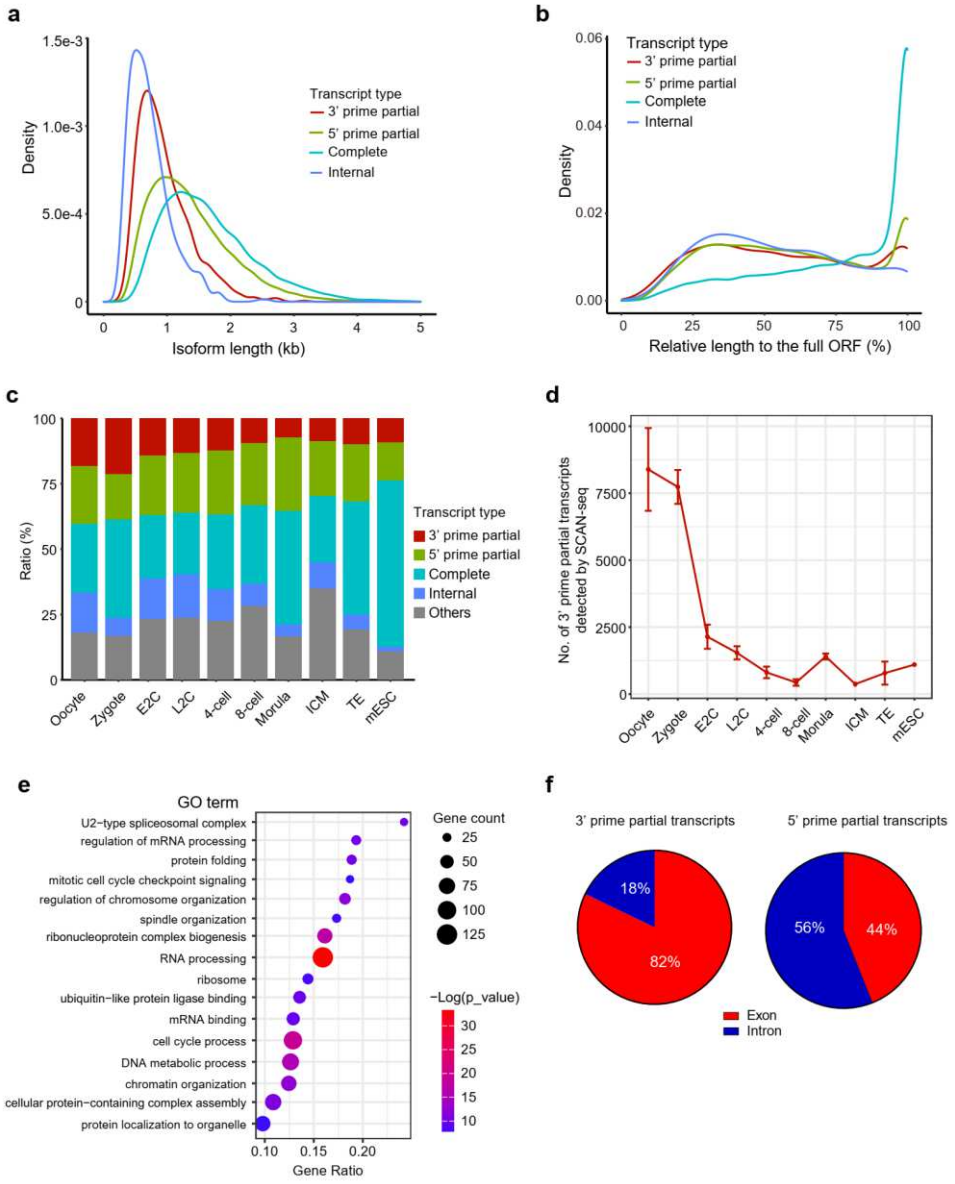
a, Diagram of the experimental and analysis workflow for single-cell isoform sequencing of the mouse preimplantation embryos. **b**, Saturation curve of representative cells from each stage. **c**, Correlation between detected UMI counts and absolute spiked abundances of each ERCC gene. **d**, Isoform mapping results of the SIRV spike-ins.



480

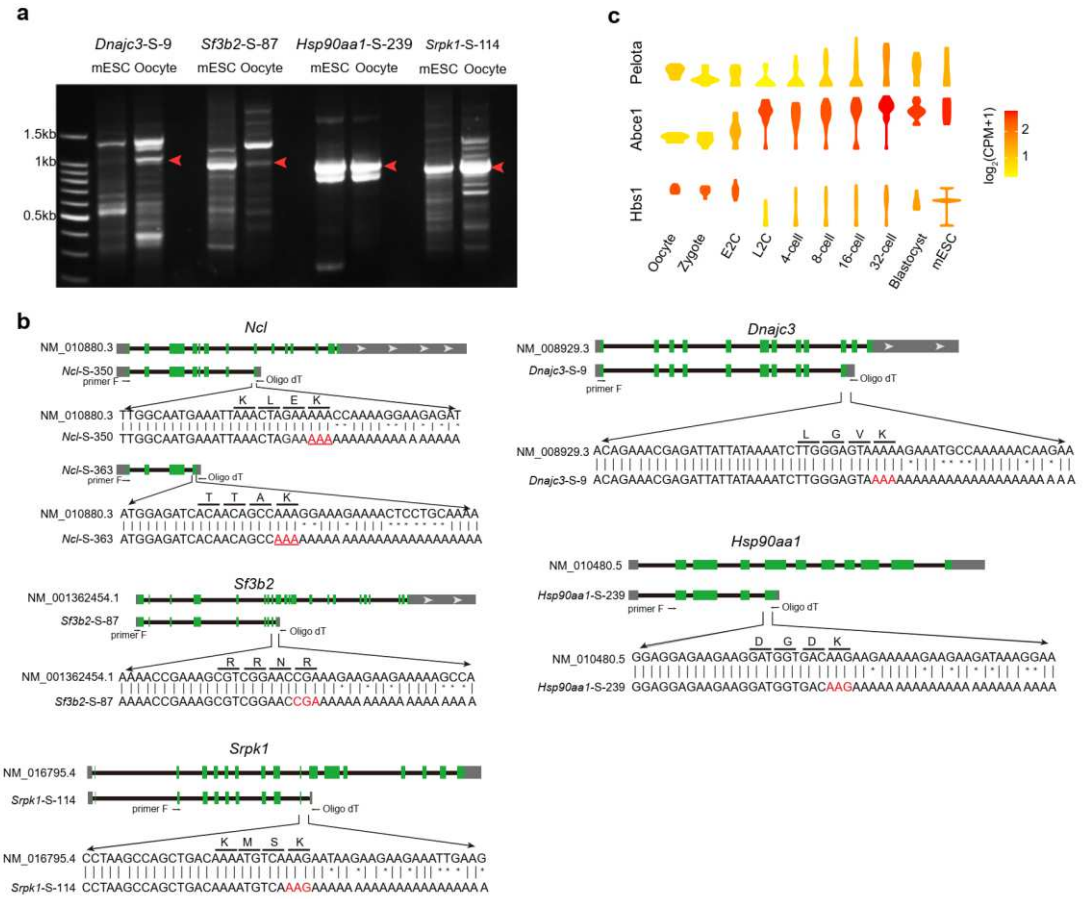
481 **Extended Data Fig. 2| Relationship between gene and isoform expression**

482 **a-b**, The ratios of genes detected with different numbers of isoform types for each stage of
483 mouse embryos and mESCs in this study (**a**) and SCAN-seq data (**b**). **c-d**, Expression
484 levels of genes detected with different numbers of isoform types for each stage of mouse
485 embryos and mESCs in this study (**c**) and SCAN-seq data (**d**). **e**, Density plot showing the
486 proportion of the major isoforms in genes expressing multiple isoform types. Only genes
487 detected with UMI counts over 5 were included.



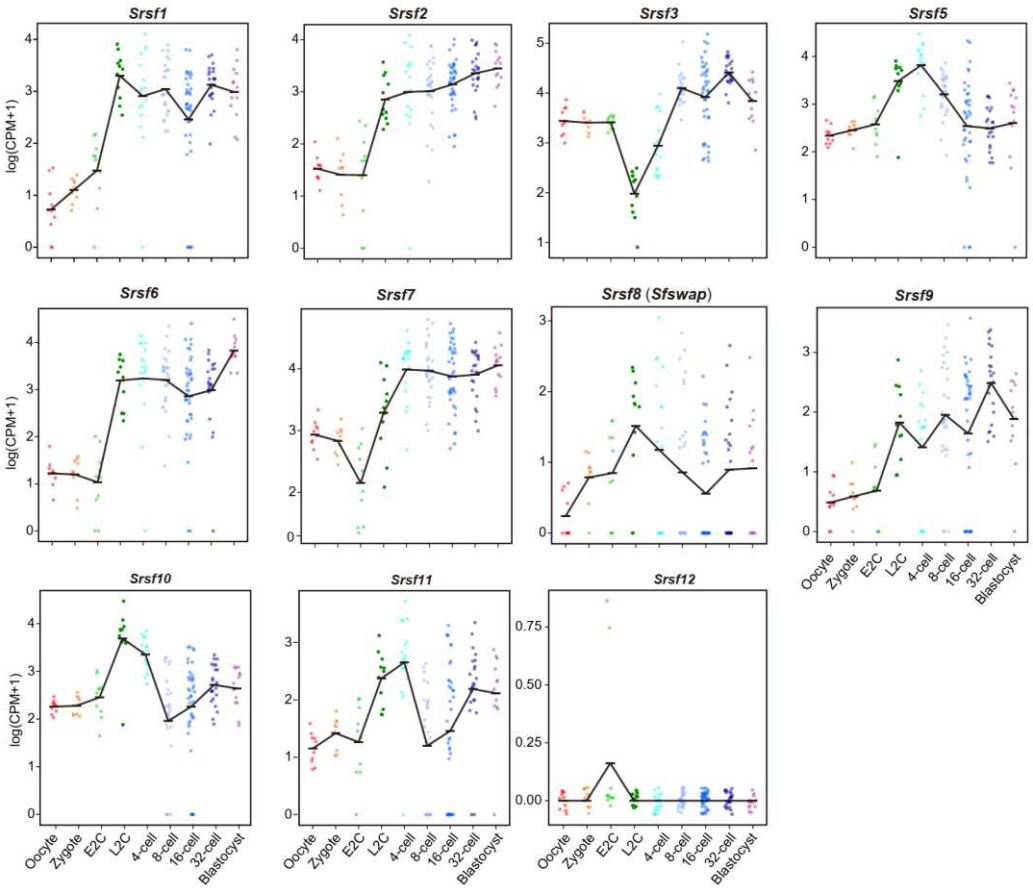
Extended Data Fig. 3| The characteristics of different types of transcripts

a, Length distribution of different types of transcripts. **b**, Relative length of the predicted protein to the complete reference ORF of each type of transcript. **c**, Ratios of each type of transcript at different stages calculated using SCAN-seq data. **d**, Expression level of the 3' prime partial transcripts detected at each stage in SCAN-seq data. **e**, The top GO terms for genes generating 3' prime partial transcripts in the mouse embryo samples. **f**, The ratios of 3' and 5' prime partial transcripts ended in exons and introns.



Extended Data Fig. 4| Validation of the 3' prime partial transcripts

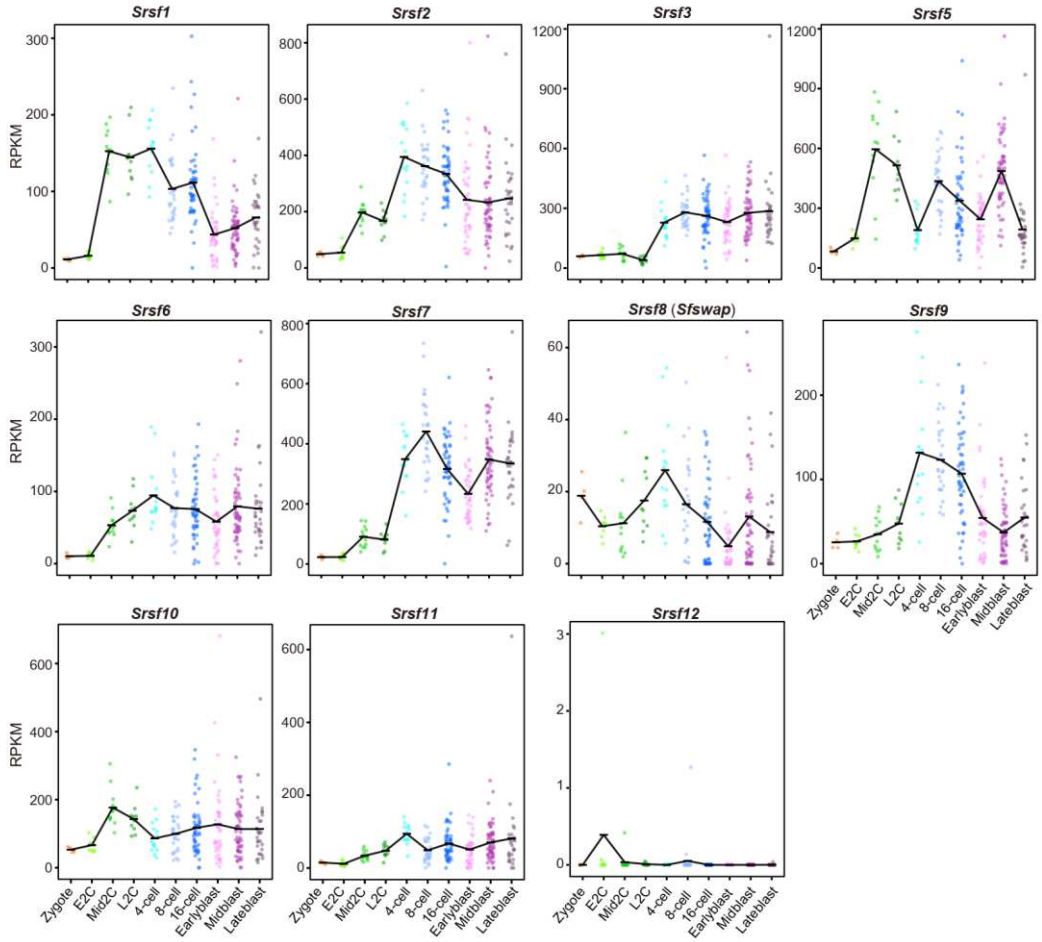
a, Gel picture showing the isoforms by RT-PCR of the target genes in mESCs and mouse oocytes. The red arrows indicate the candidate 3' prime partial transcripts of the genes. **b**, Sanger sequencing of the candidate 3' prime partial transcripts in panel **a** and the *Ncl* gene. **c**, Violin plot showing the expression levels of the essential NSD genes at different mouse embryonic stages.



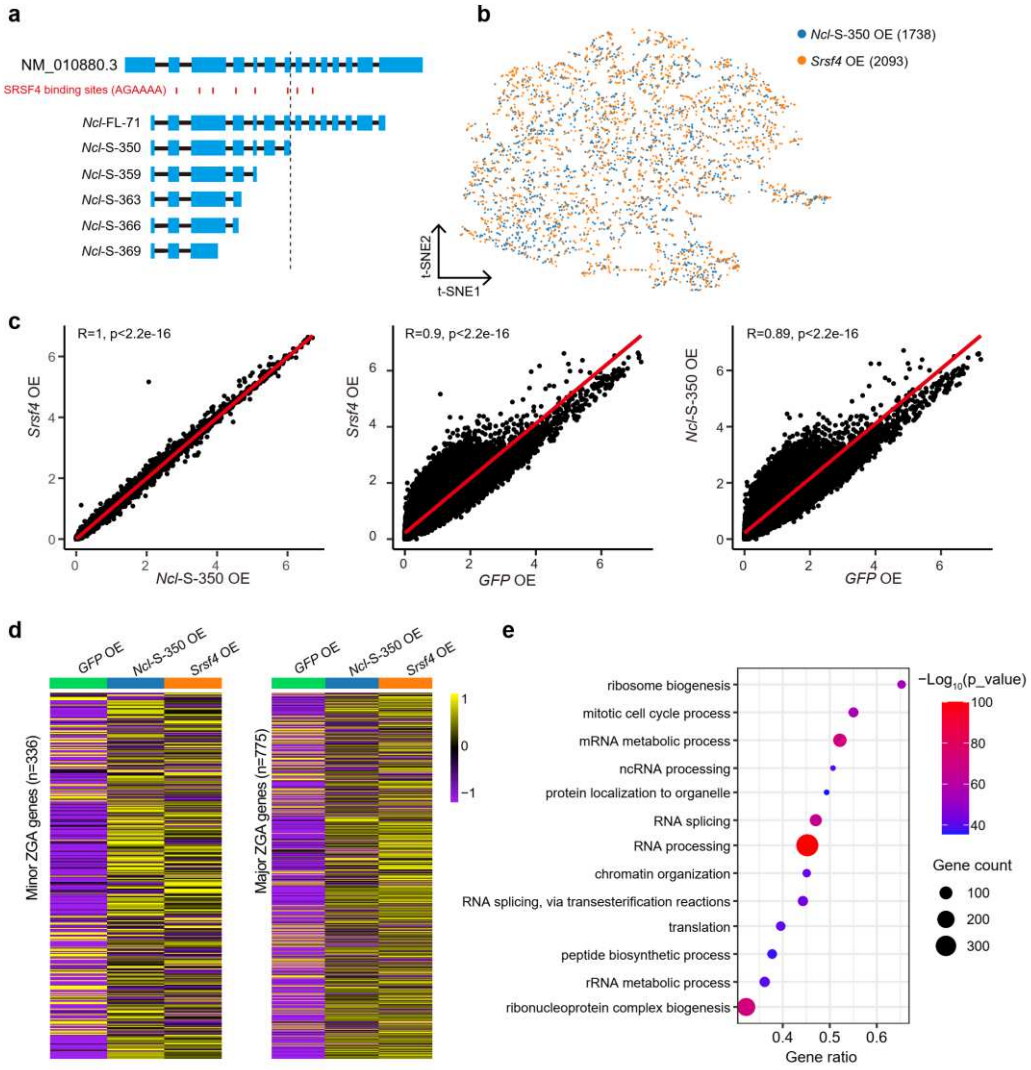
503

504 **Extended Data Fig. 5| Expression levels of each SR family gene at different mouse**

505 **embryonic stages in our data.**



Extended Data Fig. 6| Expression level of each SR family gene at different mouse embryonic stages in the NGS data by Deng *et al*[39].



Extended Data Fig. 7| Transcriptomic changes after overexpression of *Srsf4* or *Ncl*-S-350 in mESCs

a, The location distribution of the SRSF4-specific binding motif on the *Ncl* transcript. **b**, The t-SNE map showing the high consistency of the *Srsf4* OE and *Ncl*-S-350 OE mESCs based on gene expression patterns. **c**, Gene expression correlation between the *Srsf4* OE sample and the *Ncl*-S-350 OE sample, the *GFP* OE sample and the *Srsf4* OE sample, and the *GFP* OE sample and the *Ncl*-S-350 OE sample respectively. **d**, Heatmap showing generally elevated expression of minor and major ZGA genes in *Srsf4* OE and *Ncl*-S-350 OE cells compared with control *GFP* OE cells. **e**, The top GO terms for genes generating 3' prime partial transcripts in the *Srsf4* OE mESCs

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendeddataTable1.xlsx
- ExtendeddataTable2.xlsx
- ExtendeddataTable3.xlsx
- ExtendeddataTable4.xlsx
- ExtendeddataTable5.xlsx
- ExtendeddataTable6.xlsx