PLOS BIOLOGY

METHODS AND RESOURCES

Single-cell analysis of isoform switching and transposable element expression during preimplantation embryonic development

Chaoyang Wang[1,2,3☯], Zhuoxing Shi[3☯], Qingpei Huang[1,2], Rong Liu[1,2], Dan Su[1,2], Lei Chang[1,2], Chuanle Xiao[3], Xiaoying Fan[1,2,4]*

1 GMU-GIBH Joint School of Life Sciences, The Fifth Affiliated Hospital of Guangzhou Medical University, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou, China, **2** The Bioland Laboratory (GuangZhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, China, **3** State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou, China, **4** The Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China

☯ These authors contributed equally to this work.
* fan_xiaoying@gzlab.ac.cn

Citation: Wang C, Shi Z, Huang Q, Liu R, Su D, Chang L, et al. (2024) Single-cell analysis of isoform switching and transposable element expression during preimplantation embryonic development. PLoS Biol 22(2): e3002505. https://doi.org/10.1371/journal.pbio.3002505

Academic Editor: Bon-Kyoung Koo, Center for Genome Engineering, REPUBLIC OF KOREA

Received: June 9, 2023

Accepted: January 18, 2024

Published: February 16, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3002505

Data Availability Statement: All single-cell isoform sequencing data generated in this study are available at NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under the

Abstract

Alternative splicing is an essential regulatory mechanism for development and pathogenesis. Through alternative splicing one gene can encode multiple isoforms and be translated into proteins with different functions. Therefore, this diversity is an important dimension to understand the molecular mechanism governing embryo development. Isoform expression in preimplantation embryos has been extensively investigated, leading to the discovery of new isoforms. However, the dynamics of isoform switching of different types of transcripts throughout the development remains unexplored. Here, using single-cell direct isoform sequencing in over 100 single blastomeres from the mouse oocyte to blastocyst stage, we quantified isoform expression and found that 3-prime partial transcripts lacking stop codons are highly accumulated in oocytes and zygotes. These transcripts are not transcription by-products and might play a role in maternal to zygote transition (MZT) process. Long-read sequencing also enabled us to determine the expression of transposable elements (TEs) at specific loci. In this way, we identified 3,894 TE loci that exhibited dynamic changes along the preimplantation development, likely regulating the expression of adjacent genes. Our work provides novel insights into the transcriptional regulation of early embryo development.

Introduction

A gene can be transcribed into various isoforms, which are then translated into different proteins. Isoform compositions differ between cell types and states, making isoform switching a crucial factor in determining cell identity [1,2]. Third-generation sequencing-based single-cell RNA-sequencing methods like SCAN-seq, HIT-scISOseq, and MAS-ISO-seq have been developed to directly sequence gene isoforms [1,3–7]. SCAN-seq is known for its high gene

accession number GSE250381. Code availability The HIT-scISOseq analysis pipeline and source code are available from https://github.com/shizhuoxing/scISA-Tools. Additionally, the source code utilized in this study has been published at: https://zenodo.org/records/10394889.

Funding: This work was supported by grants from the National Key Research and Development Program of China (2020YFA0112201 to XF), the National Natural Science Foundation of China (32071451 to XF), the Guangdong Provincial Pearl River Talents Program (2021QN02Y747 to XF), and the R&D Program of Guangzhou National Laboratory (SRPG21-001 to XF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: CCS, circular consensus sequencing; E2C, early 2-cell; GO, Gene Ontology; hCG, human chorionic gonadotropin; L2C, Late 2-cell; LINE, long interspersed element; LTR, long terminal repeat; mESC, mouse embryonic stem cell; MZT, maternal to zygote transition; ORF, open reading frame; PCA, principal component analysis; PMSG, pregnant mare's serum gonadotropin; RRM, RNA recognition motif; SINE, short interspersed element; SRA, Sequence Read Archive; TE, transposable element; TSS, transcription start site; ZGA, zygotic genome activation.

detection sensitivity and ability to detect many novel transcripts in rare samples [4]. However, it fails to quantify the absolute abundance of genes and isoforms due to the high error rate of Nanopore sequencing [8,9]. On the other hand, HIT-scISOseq and MAS-ISO-seq use the PacBio HiFi sequencing platform to quantify isoform abundance in single cells with improved data throughput [6,7].

Isoform switch plays an important role in cell fate determination. *PBX1*, for example, can be transcribed into 3 different isoforms, each with distinct functions. *PBX1a* maintains the pluripotency of mouse embryonic stem cells (mESCs), while *PBX1b* promotes differentiation [10]. Other genes such as *Tcf3* and *Sall4* have similar regulatory patterns in mESCs [11,12]. The molecular regulation of preimplantation embryo development has been the focus of many studies, particularly maternal to zygote transition (MZT), which is crucial for whole-body development [13–16]. Although hundreds of genes have been identified in zygotic genome activation (ZGA), the functional regulators remain largely unclear, including whether isoform switching participates in the process [17–19].

Transposable elements (TEs) account for approximately 46% and approximately 37.5% of the human and mouse genome, respectively [20,21], contributing to evolution and genetic regulation. They can be divided into 2 major classes based on the transposition mode: class I retrotransposon and class II DNA transposon [22,23]. Class I, which makes up about 95% of total TEs, includes long and short interspersed elements (LINEs and SINEs, respectively) and long terminal repeats (LTRs). TEs are known to play a crucial role during embryo development [24]. For instance, MERVL and MT2_mm (a truncated form of ERVL containing only the LTR domain) can serve as promoters for totipotent gene expression, and their expression has been considered an essential totipotent biomarker [18,25,26]. LINEs, particularly LINE1, have been reported to suppress the expression of totipotent genes such as *Dux* [27,28]. A previous study showed that the hominoid-specific transposon (SINE-VNTR-Alu) acts as an enhancer to promote the ZGA process [29]. However, due to their highly repetitive nature, it is challenging to determine the activity of TEs at the locus level with limited read length. Analyzing TE expression in specific loci is therefore important for gene transcriptional regulation.

In this study, we adapted the HIT-scISOseq method for low-throughput cell analysis and sequenced isoforms in single blastomeres of mouse preimplantation embryos [6]. We analyzed cell heterogeneity within the same embryos at the same stage, providing insights into the timing of cell fate diversification during preimplantation embryo development. Isoforms of each gene in every single cell were quantified, and different isoform types showed varied proportions across embryonic stages. Notably, a significant number of 3-prime partial transcripts (lacking stop codons and generating proteins lacking C-termini) were observed in oocytes and zygotes, but were quickly degraded at the early 2-cell (E2C) stage. Furthermore, locus-specific TEs were analyzed, revealing dynamic expression changes during embryonic development. These TEs showed high correlation with the expression of adjacent genes, indicating their potential importance in developmental events.

Results

Modified HIT-scISOseq for the mouse preimplantation embryos sequencing

To identify gene isoforms in each blastomere of mouse oocytes and preimplantation embryos, we amplified RNAs in individual cells using a 10× gel bead and the Smart-seq2 protocol [6,30]. Subsequently, the amplified cDNAs from different cells were combined for ligation and PacBio library construction following the HIT-scISOseq method [6]. Concurrently, the barcode sequence of each cell was predetermined through Sanger sequencing of the cDNAs (Fig 1A).

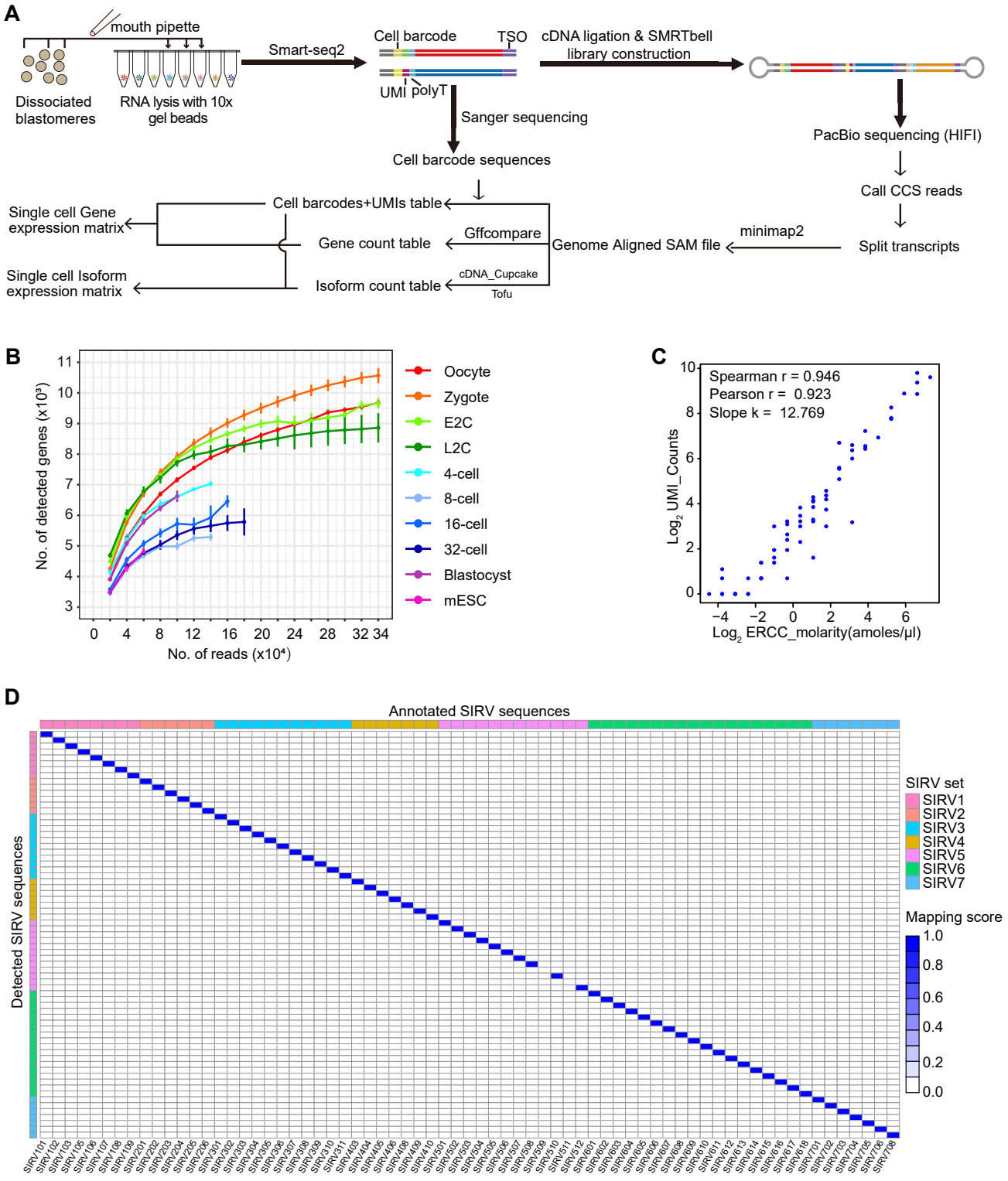

Fig 1. Quality evaluation of the single-cell isoform expression data. (A) Diagram of the experimental and analysis workflow for single-cell isoform sequencing of mouse preimplantation embryos. (B) Saturation curve of representative cells from each stage. The raw data for this plot is supplied in S1 Data. (C) Correlation between detected UMI counts and absolute spiked abundances of each ERCC gene. The raw data for this plot is supplied in S1 Data. (D) Isoform mapping results of the SIRV spike-ins. The raw data for this plot is supplied in S1 Data.

https://doi.org/10.1371/journal.pbio.3002505.g001

Data on gene and isoform expression from 161 single blastomeres were collected from 3 batches, covering various developmental stages and mESC (Table 1). Each sequencing batch produced approximately 5 million circular consensus sequencing (CCS) reads in 1 SMRT Cell 8M, with an average length of around 4 kb, indicating ligation of 2 to 3 cDNA molecules in most cases. After data splitting and mapping, about 90% of the isoforms could be accurately assigned to cells (Table 1). This approach allowed for relatively deep sequencing of samples from each stage (Fig 1B). To assess the precision of measuring absolute numbers of isoforms using our method, we also amplified ERCC and SIRV spike-ins [31,32]. At the gene level, we observed high correlation values between the added molecules and the detected UMI counts (Fig 1C). At the isoform level, different isoforms of the same SIRV gene were accurately identified without any false matches (Fig 1D). These findings demonstrate that our workflow precisely measures transcript abundance in each single cell.

Gene and isoform expression patterns in the mouse preimplantation embryos

The mouse oocyte and zygote contain a higher number of RNA molecules compared to later stage blastomeres due to maternally inherited RNA degradation [33]. The number of transcript molecules was strongly correlated (R = 0.96) with the number of expressed genes in the cells (Fig 2A). Principal component analysis (PCA) using gene expression data and isoform expression data showed that blastomeres of different stages were clearly separated (Fig 2B and 2C). The oocyte and zygote exhibited similar expression patterns; Late 2-cell (L2C) and 4-cell blastomeres were grouped together; the 8-cell, 16-cell, and 32-cell stages were similar; and the blastocyst cells were analogous to the mESCs. Stage-specific genes and transcripts were extracted, resulting in 3,867 genes and 6,819 isoforms, respectively (S1 Table). These were divided into 6 corresponding clusters based on their expression patterns across all embryonic stages (Fig 2D and 2E and S1 Table). Cluster 1 (C1) transcripts were highly abundant in oocytes and zygotes, subsequently degraded from E2C stage. Cluster 2 (C2) included transcripts that were only up-regulated in the 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). More genes were identified in each cluster at the isoform level, and most of the isoforms were consistent with the genes (Fig 2F and S1 Table). The results indicate that single-cell isoform expression data can be used to illustrate cellular heterogeneity and distinguish different types of cells as single-cell gene expression does.

Furthermore, isoforms were found to show different expression patterns compared to the host genes. Isoform switch events largely occurred during the transition from zygote to E2C and from E2C to L2C stages, during the time windows of minor and major ZGA, respectively (Fig 2H and S1 Table). For example, *Cfdp1* increased the gene expression from E2C to L2C, with its isoform *PB.73528.5* showing the same pattern, but the other isoform, *PB.73528.7*, was largely down-regulation at the same stage (Fig 2G). The expression levels of *Cnot7* gene and its isoforms *PB.71257.4* and *PB.71257.165* decreased from E2C to L2C, while *PB.71257.450*, another isoform of the gene was inversely changed (Fig 2G). Gene Ontology (GO) analysis revealed that the genes happened with isoform switch events were enriched in cytoplasmic translation, cell division, mRNA and DNA metabolic processes, etc. (Fig 2I). Functional and structural changes were identified in only 7 genes using FunFam [34,35] (S1 Fig). These results indicate that isoform switch regulation largely exists during embryonic development, especially during the ZGA process, in addition to gene expression level.

Table 1. Quality evaluation of samples from 3 batches.

	Library	Batch 1	Batch 2	Batch 3
Polymerase Reads	Polymerase Reads	5534582	6002833	7099439
	Polymerase Yield (GB)	427.02	540.46	410.25
	Polymerase Max Length	479619	493469	483234
	Polymerase Mean Length	77155.42	90033.59	57786.35
	Polymerase Read N50	154554	167339	130615
Subreads	Subread Yield (GB)	420.1	535.33	404.78
	Subreads Max Length	479619	493469	483234
	Subreads Mean Length	3001.47	4482.57	3127.49
	Subread N50	3557	5186	3508
CCS	CCS Reads	4412788	5079207	5914414
	CCS Yield (GB)	15.6	25.46	23.5
	CCS Max Length	27871	27593	27810
	CCS Mean Length	3534.33	5013.54	3973.65
	CCS N50 Length	4216	5791	4760
	CCS Mean Passes	27	19	17
	CCS MeanQV	0.96	0.97	0.94
Full-length (FL) Isoform Detection	All Paired	11184443	20967437	11008106
	FL	10013148	19427495	9427600
	Non-FL	206358	283797	283953
	Unknow	964937	1256144	1296552
	FL (%)	89.53	92.66	85.64
	Non-FL (%)	1.85	1.35	2.58
	Unknow (%)	8.63	5.99	11.78
	FL MeanLen	934.75	888.57	1081.15
	FL N50	1128	1044	1365
Cell Barcode (CB) Identification	FLNC	10013148	19427495	9427600
	CB in Whitelist	8858087	17083701	7834317
	CB in Whitelist (%)	88.46	87.94	83.1
	CB Correction	254281	553029	334615
	CB Correction (%)	2.54	2.85	3.55
	Total Corrected CB	9112368	17636730	8168932
	Total Corrected CB (%)	91	90.78	86.65

https://doi.org/10.1371/journal.pbio.3002505.t001

Isoform diversity decreases along preimplantation embryo development

To explore the connection between gene and isoform expression during mouse preimplantation development, we grouped genes into 6 categories based on the number of isoform types they expressed (S2A Fig). While most genes expressed only 1 type of isoform across different stages, more genes expressed multiple types of isoforms in the earlier stages. In mouse oocytes and zygotes, around 60% of genes expressed more than 1 type of isoform, and nearly 20% of genes were found with over 5 types of isoforms. In contrast, approximately 70% of genes in mESCs expressed only 1 type of isoform, and less than 5% of genes expressed more than 5 types of isoforms (S2A Fig). The same isoform expression characteristics were observed in SCAN-seq data (S2B Fig), indicating a diverse range of isoforms in early mouse embryos [4]. To rule out the possibility that this observation was caused by higher mRNA abundance in early embryos, especially in oocytes and zygotes, we performed oocyte splits. The results showed that the ratios of genes containing different numbers of isoform types were almost consistent among intact oocytes, 1/2 oocytes, and 1/4 oocytes (S2C Fig), suggesting that the

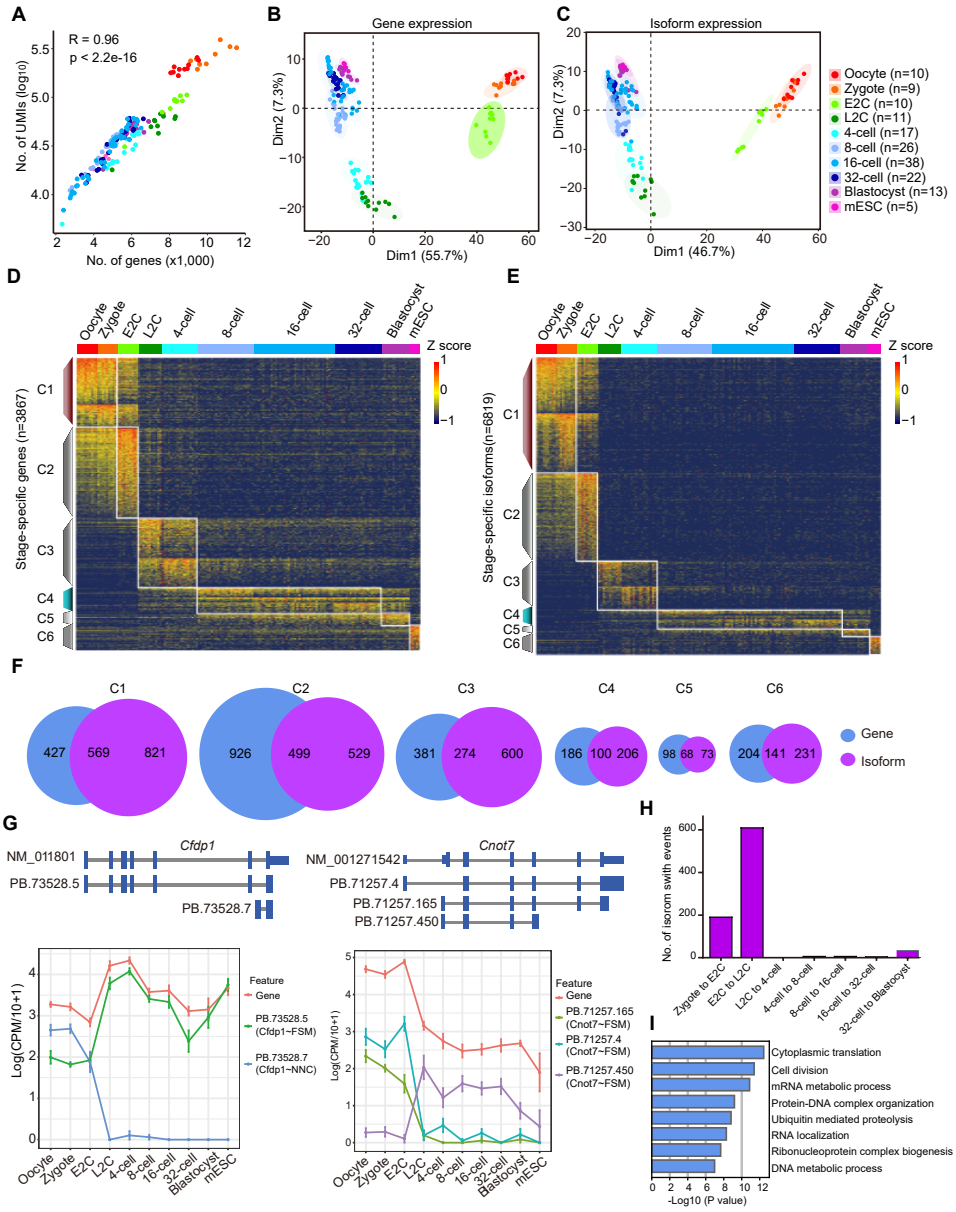

Fig 2. Gene and isoform expression in the mouse preimplantation embryos. (A) Number of genes and UMIs detected in each cell at different stages. The raw data for this plot is supplied in S2 Data. (B, C) PCA plot of all the blastomeres and mESCs based on gene expression (B) and isoform expression (C). The PCA loadings are supplied in S1 Table. The raw data for these 2 plots are supplied in S2 Data. (D, E) Heatmap of stage-specific genes (D) and isoforms (E). The raw data for these 2 plots are supplied in S2 Data. (F) Venn plot of pairwise groups of stage-specific genes and isoform-corresponding genes. (G) Isoform switch during embryo development, the upper picture is the reference transcript and our sequence isoform and the lower part is the gene and main isoform expression of *Cfdp1* and *Cnot7*. The raw data for this plot is supplied in S2 Data. (H) Number of isoform switch events between each adjacent embryonic stages. The raw data for this plot is supplied in S2 Data. (I) GO results of genes showed isoform switch during preimplantation embryo development. The raw data for this plot is supplied in S2 Data. GO, Gene Ontology; mESC, mouse embryonic stem cell; PCA, principal component analysis.

https://doi.org/10.1371/journal.pbio.3002505.g002

detected isoform diversity was hardly affected by the amount of mRNAs. Additionally, the genes expressing more types of isoforms were detected with higher expression levels in both our data and SCAN-seq data (S2D and S2E Fig). To validate this hypothesis, we randomly selected 3 highly expressed genes (CPM $>$ 100) and 3 lowly expressed genes (CPM $<$ 10) in mESC to confirm their isoform diversity by reverse transcription and PCR (RT-PCR). Although there were more types of isoforms revealed by RT-PCR than the sequencing results, the highly expressed genes still showed higher isoform diversity (S2F Fig). We then assessed the isoform dominant level in each gene expressing multiple types of isoforms by calculating 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 (S2G Fig). 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.

Large abundance of 3-prime partial transcripts are observed in mouse oocytes and zygotes

Based on the putative integrity of corresponding open reading frames (ORFs), the transcripts were categorized into 5 types: complete isoforms encoding the full ORFs, 3-prime partial transcripts and 5-prime partial transcripts lacking the stop codon and start codon sections respectively, internal transcripts predicted with proteins lacking both ends, and others where the detected ORF lengths in the isoforms were below the software-set threshold [36] (Fig 3A). When compared to the annotated transcription start site (TSS), the 5-prime partial transcripts exhibited the lowest overlap ratio with the CAGE peaks (S3A Fig), indicating that some of these transcripts might be generated by incomplete reverse transcription.

As anticipated, the complete transcripts displayed the longest lengths, while the internal transcripts were the shortest (S3B Fig). However, the predicted protein length was similar for the 3 incomplete transcript types (S3C Fig). Intriguingly, we observed that the 3-prime partial transcripts were highly expressed in oocyte and zygote, but their expression dramatically decreased from the E2C stage (Fig 3B and 3C). This expression pattern was also observed in the SCAN-seq data (S3D and S3E Fig). Subsequently, we conducted GO analysis on genes detected with 3-prime partial transcripts (S2 Table). These genes were enriched in pathways related to RNA processing, cell cycle checkpoint, ribonucleoprotein complex biogenesis, DNA metabolic process, chromatin organization, etc. (Fig 3D), all of which are known to play essential roles in mouse and human preimplantation embryo development [13,16,37–40].

We then selected some candidates to validate the enrichment of 3-prime partial transcripts. The host genes related to RNA processing (*Sf3b2*, *Srpk1*) and protein translation and transporting (*Dnajc3*, *Hsp90aa1*) were revealed by gel analysis of mouse oocyte RT-PCR products (Fig 3E). Furthermore, no stop codons were identified in these transcripts according to the Sanger sequencing results (Fig 3F).

Expression of *Ncl* showed significantly isoform switch during embryonic development

The Nucleolin gene encodes NCL, which is involved in various cellular processes such as ribosome biogenesis, chromatin organization and stability, and DNA and RNA metabolism [41]. It also regulates totipotent genes expression with KAP1 and LINE1 [27,28]. Isoform sequencing revealed that the *Ncl* gene encodes 6 isoform types, categorized based on the number of RNA recognition motifs (RRMs) they contain (Fig 4A). RT-PCR showed that the short isoform (*Ncl-S-350*) was more abundant than the complete isoform (*Ncl-FL-71*) in mouse oocyte

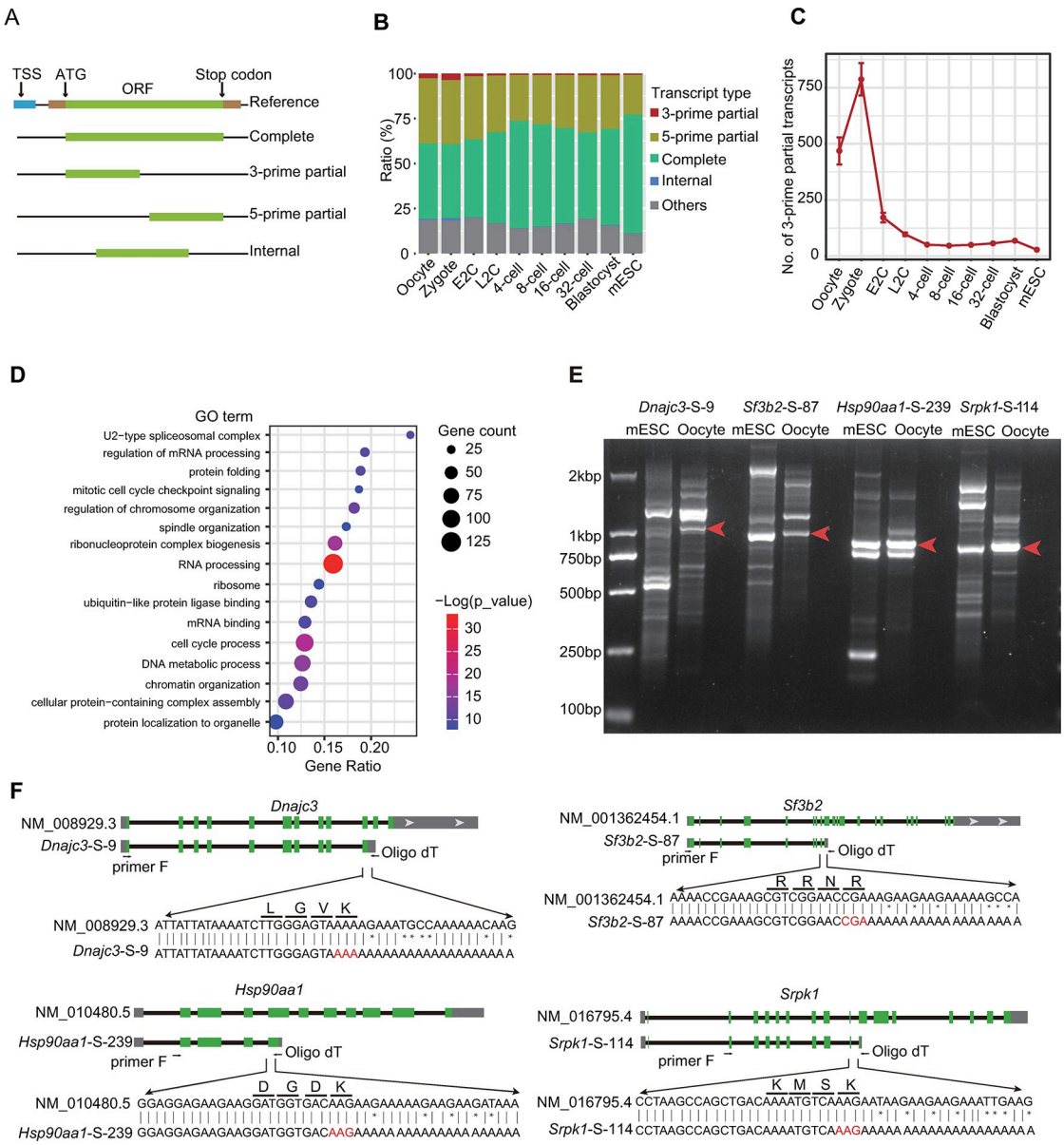

Fig 3. Expression patterns of different types of transcripts during mouse preimplantation embryo development. (A) Schematic diagram of different transcript types. (B) Ratios of each type of transcripts at different stages. The raw data for this plot is supplied in S3 Data. (C) Number of 3-prime partial transcripts detected at each stage. The raw data for this plot is supplied in S3 Data. (D) The GO analysis of 3-prime partial transcripts. The raw data for this plot is supplied in S3 Data. (E) Gel picture showing the isoforms by RT-PCR of *Dnajc3*, *Sf3b2*, *Hsp90aa1*, and *Sprk1* in mouse oocytes and mESCs. The raw image for this plot is supplied in S1 Raw Images. (F) Sanger sequencing of the candidate 3-prime partial transcripts in Fig 3E. GO, Gene Ontology; mESC, mouse embryonic stem cell.

https://doi.org/10.1371/journal.pbio.3002505.g003

(Fig 4B). The 2 of most enriched short isoforms were confirmed as 3-prime partial isoform by Sanger sequencing (Fig 4C). *Ncl* abundance was first down-regulated in the E2C stage and then increased from the L2C stage at the gene level (Fig 4D). The 2 categories of short *Ncl* isoforms were highly expressed in oocytes and zygotes and then almost disappeared. Conversely, the complete *Ncl* isoform showed lower expression in maternal RNA and was largely up-regulated after ZGA (Fig 4D). Our findings highlight the isoform switch of the *Ncl* gene during the ZGA process, which is masked in gene-level analysis.

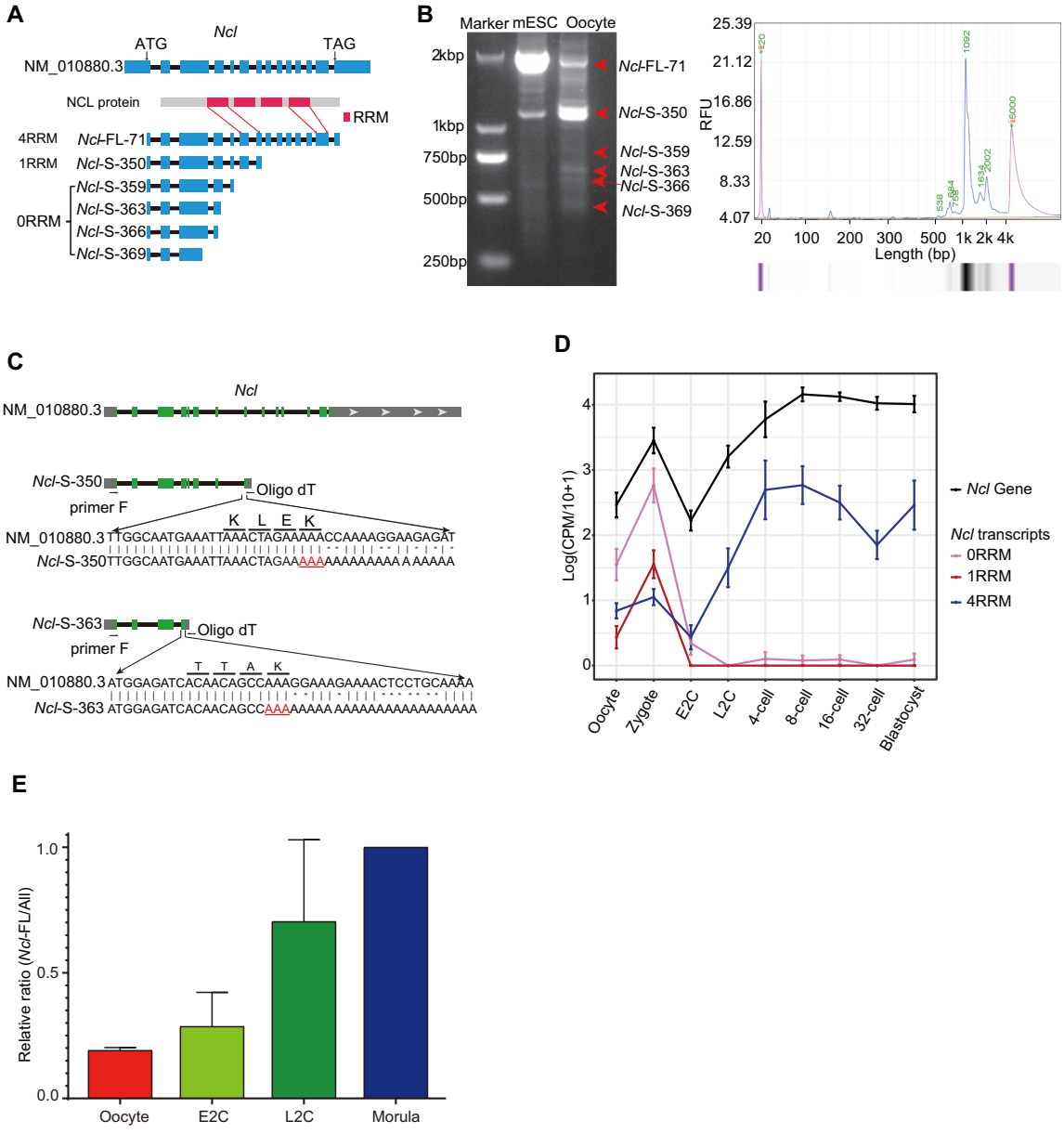

Fig 4. Isoform expression pattern of *Ncl* during mouse preimplantation embryonic development. (A) Schematic diagram of *Ncl* isoforms. (B) Gel picture (left) and Q-sep result (right) of *Ncl* RT-PCR products from mouse oocyte. The raw image for this plot is supplied in S1 Raw Images. (C) The Sanger sequencing results of the top 2 enriched short *Ncl* isoforms. (D) Expression levels of each category of *Ncl* isoforms and the *Ncl* gene at different stages. The raw data for this plot is supplied in S4 Data. (E) The relative ratios of *Ncl* full-length isoforms at different stages detected by RT-qPCR. The relative ratio of *Ncl*-FL/All in morula was set as 1.0. The morula had 1 sample, while other stages all had 3 replicates. The raw data for this plot is supplied in S4 Data.

https://doi.org/10.1371/journal.pbio.3002505.g004

To demonstrate this result, we performed reverse transcription and real-time quantitative PCR (RT–qPCR) using primers targeting all the *Ncl* isoform types or only the complete 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 are complete in oocytes when we set the full-length relative ratio as 100% at the morula stage (Fig 4E). This result confirmed the dynamic isoform switch during embryo development in vivo.

TEs are dynamically activated during embryonic development

Due to the repetitive and interspersed features of TE sequences and their transcripts, TGS-based sequencing is more suitable for TE research [42,43]. Our long-read and highly accurate results can help us investigate TE expression at specific loci. TE expression was quantified in each single cell. Generally, the amount of TE RNAs belonging to different super-families decreased along embryonic stages (Fig 5A). Although maternal RNA contains the largest pool of TE elements, zygotes were detected with more TE RNA copies, suggesting TE as an important regulator to promote minor ZGA (Fig 5A). Additionally, TE expression elevated from 32-cell to blastocyst stage, indicating that TEs play a role in embryonic pluripotent stem cells. Our single-cell direct isoform sequencing data enabled mapping the TE reads to specific loci confidently, and we also calculated the number of expressed TE loci at each stage. Hundreds of TE loci were transiently transcribed at the zygote stage, further supporting the deduction that TEs regulate minor ZGA (Fig 5B). More active TE loci were detected in blastocysts than morulae, also indicating the important role of TEs in embryonic pluripotent stem cells (Fig 5B).

We further calculated the TE expression level according to total reads or single locus mapped reads belonging to each TE superfamily in each single cell (Fig 5C). In both calculation ways, the oocyte and zygote were detected with higher expression levels. However, more different expression patterns were observed between total TE superfamily expression and individual TE locus expression. For example, LINE was slightly down-regulated from E2C to 32-cell stage when looking at the superfamily, but the transcription level at each locus was up-regulated. Differently, the total LTR expression transiently increased at L2C stage, but each locus was detected with lower expression level (Fig 5C). These differences still exist when calculating in each TE family (S4A and S4B Fig). The ERVL, which has been proven to be associated with totipotent genes' activation [18,25,26], is indeed the highest expressed in L2C samples. However, each active ERVL site in the L2C genome did not express the most copies of corresponding RNAs. This indicates that more ERVL sites are transiently active to regulate a large scale of major ZGA genes at L2C stage (S4A and S4B Fig). Gaining information on locus-specific TE expression may help us to gain a deeper understanding of how TEs regulate different developmental processes.

To study the role of specific TEs at distinct locations, not categorized in a family or subfamily, in regulating preimplantation embryo development, we sought out TE loci with stage-specific expression patterns. We identified a total of 3,894 TE loci, which could be classified into 5 clusters based on their expression patterns across all embryonic stages (Fig 5D and S3 Table). Specifically, Cluster 1 (C1) TEs exhibited higher expression in oocytes and zygotes. Cluster 2 (C2) TEs were predominantly expressed in the E2C stage. Cluster 3 (C3) and Cluster 4 (C4) TEs showed high expression in the L2C to 4-cell stages and 8-cell to blastocyst stages, respectively. The mESC-specific TEs were grouped in Cluster 5 (C5).

Subsequently, we explored the involvement of ERVL and LINE1 subfamilies in the MZT process and later stage development [18,25–27] (Figs 5E, 5F, S4C and S4D). As anticipated, MERVL-int and MT2_mm, which only contain the LTR promoter of MERVL element, were the primary active MERVL subtypes in C3, a stage during which major totipotent genes are expressed [19,44] (Fig 5E). Conversely, MT2B1 was the most active subtype in the maternal genome. The LINE1 superfamily has been reported to silence totipotent genes such as *Dux* [27]. In our data, Lx7 emerged as the most active LINE1 subfamily since the L2C stage (Figs 5F and S4D).

A TE subfamily consists of hundreds to thousands of TE copies from different loci, and these copies are transcribed independently. We observed diverse activation of different TE copies even within the same subfamily (S5A and S5B Fig). For example, although MERVL-int and MT2_mm participate in ZGA, some copies from chromosome 1 and chromosome 5 were

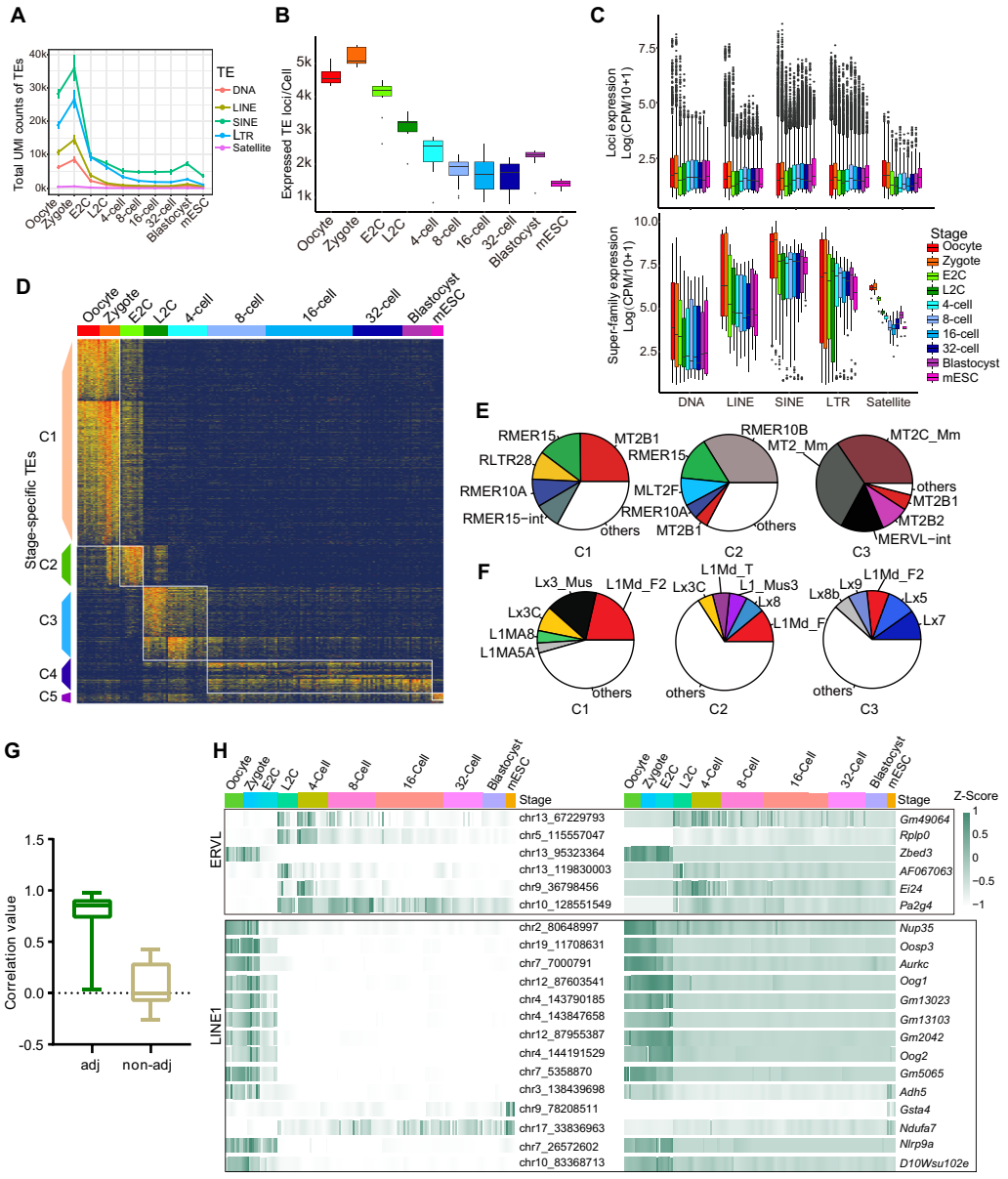

Fig 5. Dynamic expression patterns of TEs during mouse preimplantation embryo development. (A) UMI counts of each TE superfamily detected in single cells at each stage. The raw data for this plot is supplied in S5 Data. (B) Number of expressed TE loci per cell at each stage. The raw data for this plot is supplied in S5 Data. (C) Expression level of each TE locus (upper panel) and all TE counts (lower panel) belonging to different classes in each stage, satellite was also included. Each dot represents a cell. Only active loci in each cell were calculated. The raw data for this plot is supplied in S5 Data. (D) Expression heatmap of stage specific TEs (n = 3,894 unique loci). The TEs were ordered by their averaged expression levels within each stage (also see S3 Table). The raw data for this plot is supplied in S5 Data. (E) The top5 subfamilies of ERVL expressed in clusters C1–C3 in Fig 5D. (F) The top5 subfamilies of LINE1 expressed in clusters C1–C3 in Fig 5D. (G) Correlation values between expression levels of representative stage-specific LINE1 and ERVL loci (averaged UMI counts >5) and their adjacent genes (<10 kb distance). Non-adjacent genes are defined as genes over 1 Mb away from the TE loci. The raw data for this plot is supplied in S5 Data. (H) Heatmap showing examples of consistent expressions between specific TE loci and their adjacent genes along developmental stages. The raw data for this plot is supplied in S5 Data. LINE, long interspersed element; TE, transposable element.

https://doi.org/10.1371/journal.pbio.3002505.g005

exclusively active in the maternal genome (S5A Fig). Most of the LINE1 copies belonging to different subfamilies were actively transcribed in oocytes, but more Lx7 copies became active only after the E2C stage (S5B Fig). The expression level of TE loci showed a notably higher correlation with that of adjacent genes (Fig 5G). We observed consistent expression of specific LINE1 and ERVL loci with adjacent genes, indicating common active regulations for different TE families (Fig 5H). As TEs can function as transcriptional regulatory elements such as enhancers, detecting locus-specific TE expression provides more insights into understanding the regulatory mechanism of preimplantation embryo development [45].

Discussion

Single-cell sequencing has significantly advanced our research in preimplantation embryo development. We have identified genes, isoforms, and TE elements that are specific to different developmental stages, revealing a rich diversity of isoforms in the maternal contents. By modifying the HIT-scISO-seq to a low-throughput method, we were able to track cells from the same embryo during the experiment. This single-cell approach not only illustrated differences across developmental stages but also highlighted cell heterogeneity within a specific stage, allowing for the evaluation of fate differentiation of blastomeres within an embryo.

The 2 blastomeres in the 2-cell embryo showed high consistency in both gene and isoform expression levels, but became increasingly different from each other from the 4-cell stage (S5C and S5D Fig), that fate divergence could be revealed at the 4-cell stage on the transcriptional level. When comparing the heterogeneity between cells from different embryos at the same stage, the correlation values also decreased along with developmental stages. The decreased cell–cell correlation at the zygote stage likely resulted from batch effects of different embryo replicates. However, the low-throughput approach, affected by the low-throughput and high cost of TGS, also makes it difficult to obtain a large sample size, especially since the cells collected in the blastocyst stage are insufficient to illustrate intra-embryo heterogeneity. We believe that future studies using high-throughput approaches are more powerful in fully specifying the transcript regulations in cell fate specification at the blastocyst stage.

TGS-based isoform sequencing, at the bulk or single-cell level, has annotated an abundance of novel transcripts and splicing events in preimplantation embryos [4,17,22]. 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 a large number of 3-prime partial transcripts, which lack stop codons, in mouse MII oocyte and zygote (Fig 3B, 3C and 3E). This type of transcript has been extensively studied in cancer and is considered an oncogenic factor [46]. In early embryos, these transcripts might be important for the MZT process, as the host genes are highly enriched in biological processes responsible for mouse and human preimplantation embryo development (Fig 3D). Further studies are needed to resolve the generation and function of these transcripts.

TEs are the main components of the mammalian genome. However, their biological function is still largely unclear, and most of them were previously regarded as parasites or "junk DNA" [27]. Although some TE classes have been investigated in preimplantation embryos, previous studies almost exclusively used NGS-based analysis, only revealing the TEs at the subfamily level [22,27,47,48]. Our long-read direct isoform sequencing directly quantifies the TE transcription from different loci (Figs 5 and S3). Our data displays more detailed TE expression dynamics, which helps us to investigate these genome "dark matters" in more detail.

TEs had been reported to regulate gene expression by different ways. For example, LINE1 has the ability to increase the chromatin accessibility [21,27,49], while MERVL and MT2_mm can derive totipotent gene expression in 2-cell and 4-cell embryo [18,25,26]. Therefore,

exploring how dynamic expression of locus-specific TEs regulates gene expression requires further investigation. On the other hand, abnormal expression of TEs in most differentiated tissues is harmful to humans. For instance, LINE1 overexpression is highly related to cancers such as gastric cancer and lung squamous cell carcinoma [50–52]. HERVK overexpression is related to aging in mouse, monkey, and human [53]. Therefore, measuring TE expression at the locus-specific level offers a new way to decode the mechanisms in various human diseases.

Beyond just TE expression, several studies have identified transcript isoforms where TEs are used as alternative promoters for gene expression [18,54–57]. We also attempted to find TE chimeric transcripts in the mouse preimplantation embryos. A total of 6,143 TE chimeric transcripts were identified, with over 80% only detected in 1 cell with 1 copy (S4 Table). One reason could be the low expression levels of these transcripts. Considering the limited sequencing depth from the TGS in this study, it is difficult to fully detect these transcripts. Another reason might be the current bioinformatic methods, which are not suitable for analyzing the TE chimeric transcripts. As the recent study by Berrens and colleagues [22] also captured quite a limited number of TE-derived isoforms in each cell (approximately 200 for the mouse 2-cell sample and approximately 20 for human iPSCs). The increase of TGS throughput and the development of bioinformatics would help us to further explore on such interesting regulations.

Materials and methods

Ethics statement

All animal experiments were performed according to the guidelines of the Institutional Animal Care and the Ethics Committee of the Guangzhou Institutes of Biomedicine and Health (Guangzhou, China). The research license number is IACUC2020113.

Animals and single blastomere collection

We used 6- to 8-week-old C57BL/6J female mice and DBA/2NCrl male mice in the experiment. The female mice were first injected with 7.5 IU of pregnant mare's serum gonadotropin (PMSG) (Ningbo SanSheng Biological Technology, Cat. 110044564) and with 7.5 IU of human chorionic gonadotropin (hCG) (Ningbo SanSheng Biological Technology, Cat. 50030248) after 46 to 48 h injected. After mating, the embryos of each stage were collected at defined time periods after hCG administration [58]: 20 h (MII oocyte, no mating), 22 to 24 h (zygote), 30 to 32 h (early 2cell), 46 to 48 h (late 2cell), 54 to 56 h (4cell), 68 to 70 h (8-cell), 78 to 80 h (16cell to 32cell), and 88 to 90 h (early blastocyst). All animal experiments were performed according to the guidelines of the Guangzhou Institutes of Biomedicine and Health (Guangzhou, China). Collection of single blastomeres at each stage was carried out as previously described [4].

Single-cell cDNA amplification and TGS library construction for PacBio sequencing

We used the same amplification procedure as SCAN-seq [4], except for changing the reverse transcription primer with a 10× gel bead for each reaction for the embryonic samples. Then, each pre-amplification product was purified by 0.6× Ampure XP beads (Beckman, Cat. A63882). The concentration was measured using Qubit dsDNA HS and BR Assay Kits (Invitrogen, Cat. Q32854). The PCR product from about 60 blastomeres which were confirmed of effective amplification were pooled together in proportion to the number of amplified cycles. We took 100 ng of the pooled cDNA to build PacBio sequencing library following the protocol of HIT-scISOseq [6] and sequenced for 1 cell with HiFi mode.

Cell barcode sequence identification by Sanger sequencing

About 2 ng of the pre-amplified cDNA of each cell was further amplified using 2 × Taq Plus Master Mix (Vazyme, Cat. P212), and then cloned into T vector (Transgen, Cat. CT111-01). Next, the ligated plasmid transferred into *Trans5α* chemically competent cell (Transgen, Cat. CD201-01) by heat shock. The M13 primer were used to identify positive clones inserted with cDNA fragments. Single clones of bacteria were collected for Sanger sequencing to identify the barcode sequence of each cell.

Mouse ES cell culture

Mouse E14 Tg2A (E14) ES cells (male) were used for all experiments. The mESCs were cultured on 0.1% gelatin-coated plates in ES-FBS culture medium as previously described [27].

Validation of the 3-prime partial transcripts

The oocytes or the mESCs RNA were reverse transcription using oligo-dT primer (AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTT). Then, the anchor sequence of oligo-dT primer and a primer located in the start codon of the interested gene (*Ncl*: ATGGTGAAGCTCGCAAAGGC; *Dnajc3*: ATGGTGGCCCCCGGCTCGGTG; *Sf3b2*: ATGGCGGCGGAGCATCCCGAACCT; *Hsp90aa1*: ATGCCTGAGGAAACCCAGACCCA; *Srpk1*: ATGGAGCGGAAAGTGCTCGCGCT) were used to amplify all isoforms containing the 5′ sites. The PCR products were first checked on 1.5% agarose gel. The candidate gel bands were recovered and the sequences were confirmed by Sanger sequencing.

Validation isoform diversity of highly and lower expressed genes

The mESCs RNA were reverse transcription using oligo-dT primer as previous mentioned. Then, the forward primer that could distinguish most diverse isoforms was used to amplification target genes (*Srsf7*: ATGTCACGCTACGGGCGGTA; *Rps5*: CTGTCTGTATCAGGGCGGCG; *Rps19*: TTTCCCCTGGCTGGCAGCGC; *Plp2*: ATGGCGGATTCTGAGCGTCT; *Mrps6*: ATGCCCCGCTACGAGTTGGC; *Ss18l2*: ATGTCTGTCATCTTCGCTCCTG). All the reverse primer used was oligo-dT. The PCR products were checked on 1.5% agarose gel.

Single-cell isoform sequencing data processing

We used stand-alone versions of SMRT-Link (version 8.0.0.80529) software package to transform raw Subreads to Calling Circular Consensus Sequencing (CCS) reads with the following parameters: "—min-passes 0—min-length 50—max-length 21000—min-rq 0.75." After CCS calling, we used HIT-scISOseq analysis software kit scISA-Tools (https://github.com/shizhuoxing/scISA-Tools) for Full-Length Non-Concatemer (FLNC) reads identification, cell barcode and UMI extraction and correction.

Alignment and generation of single-cell gene expression matrix

After trimming the primers, cell barcodes, UMIs, and polyA tails, the remaining FLNC sequences were aligned to mouse genome (10x Genomics pre-build mouse mm10 reference dataset: refdata-gex-mm10-2020-A) using minimap2 (version 2.17-r974-dirty) in spliced alignment mode with the following parameters: "-ax splice -uf—secondary = no -C5." Then, we used gffcompare (version 0.11.6) to assign the mapped FLNCs to mm10 annotation gene models (10x Genomics pre-build mouse mm10 reference dataset: refdata-gex-mm10-2020-A) base on FLNCs genome alignment SAM file. Next, based on the identified cell barcodes, we

used scGene_matrix utility of scISA-Tools to generate the single-cell gene expression matrix. The expression values were normalized as copy number per 100,000 mapped reads (CPM/10).

SIRV data evaluation

The FLNC reads were aligned to the SIRVome using minimap2 (version 2.17-r974-dirty) with the following parameters: "-ax splice -uf—MD—sam-hit-only." We only annotated the reads with assigned barcodes and valid UMIs. Then, we used gffcompare (version 0.11.6) to assign the mapped FLNCs to SIRVome annotation GTF (SIRV-Set4) base on FLNCs SIRVome alignment SAM file. A confusion matrix was generated with the counts of FLNCs assigned to the primary SIRV isoforms or not using an in-house script.

Nonredundant isoforms classification and quality assessment

First, the "collapse_isoforms_by_sam.py" python script in cDNA_Cupcake software package (https://github.com/Magdoll/cDNA_Cupcake) was used to collapse mapped FLNCs to nonredundant isoforms with parameters: "—dun-merge-5-shorter." After that, we used SQANTI3 (https://github.com/ConesaLab/SQANTI3). To assess whether transcripts are within known TSSs, we aligned them using the CAGE peak data (mouse.refTSS_v3.1.mm10.bed) provided with the "—CAGE_peak" parameter in SQANTI3. We further used SQANTI3 "RulesFilter" script to filtered artifact isoforms. Isoforms which classified as FSM, ISM, NIC, and NNC were kept for downstream analysis.

Isoform type classification by ORF prediction

After being processed with SQANTI3, the FASTA file of mapped genome sequences were extracted according to the SQANTI3 output GTF file. Base on the FASTA file, we used TransDecoder (v5.5.0) for ORF extraction and prediction. For those predicted with multiple ORFs, the longest ones were selected as the representative ORF. TransDecoder assigns each detected isoform as one of 4 types based on whether then contains the start and stop codon of the reference ORF: complete, 5-prime partial, 3-prime partial, and internal. Additionally, we assigned the isoforms that did not mapped to an ORF region by TransDecoder as others.

Generation of single-cell isoform expression matrix

After the SQANTI3 procedure, the scIsoform_matrix utility of scISA-Tools was used to generate single-cell isoform expression matrix based on the identified cell barcodes. We further filtered isoforms detected in less than 5 cells and finally 68,012 isoforms in the mouse embryonic samples were preserved.

PCA analysis based on gene and isoform expression

Before PCA dimensionality reduction, we used "FindVariableGenes()" function in Seurat R package to select the top 1,000 highly variable genes and isoforms, respectively. Then, the "PCA()" function in FactoMineR was used for dimension reduction process and we used the function "fviz_pca_ind()" in factoextra R package to plot the PCA map.

Stage-specific genes and isoforms

Based on the gene expression matrix and isoform expression matrix, respectively, we used "edgeR" to find differentially expressed genes/transcripts between each pair of adjacent embryonic stages under the criterion of logFC>1 and p-value <0.01. A total of 3,867 and 6,819 stage-specific genes and transcripts were identified, respectively. These genes and transcripts

were clustered into 6 groups according to their expression patterns across all stages. Visualization of these genes' and transcripts' expression was done using R package "pheatmap."

SCAN-seq data processing

We downloaded the SCAN-seq data available from the Sequence Read Archive (SRA) database (accession number: PRJNA616184). Following the described data processing steps of SCAN-seq. Briefly, nanoplexer (https://github.com/hanyue36/nanoplexer/) was used to demultiplex barcode for each cell in the library, and nanofilt (v2.5.0) was used for filtering low-quality reads (qscore <7) and short reads (length <100 bp), then Pychopper (v2.3) (https://github.com/nanoporetech/pychopper) was used to extract full-length reads.

After obtaining the full-length reads, we generated the gene expression matrix and isoform expression matrix using the same procedure as we did for HIT-scISOseq data.

Isoform switch analysis

Based on the isoform expression matrix, we used "IsoformSwitchAnalyzeR" to identify switch isoforms between each pair of adjacent embryonic stages under the criterion of isoform_switch_q_value<0.05 and gene_switch_q_value<0.01.

TE expression analysis

To quantify TE at the locus level, we first aligned all FLNC (full-length non-chimeric) reads to the mm10 genome. Based on the TE annotation file obtained from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/rmsk.txt.gz), we calculated the overlap between each uniquely aligned FLNC read and TE loci. Subsequently, we applied filtering criteria: the starting or ending position of FLNC alignment to the genome must fall within an overlapping TE locus, and only the TE locus with the longest overlap length was considered for quantitative counting of the same FLNC.

Following the above steps, we performed aggregate counting based on the cell barcode sequences corresponding to each FLNC, enabling us to obtain quantitative expression measurements of TE loci at the single-cell level.

To quantify the expression of TE-associated chimeric transcripts, we developed an in-house script. Firstly, this script extracts the chimeric alignments from FLNC. Secondly, it employs a method to identify TE sites that overlap with FLNC, as described in the unique TE mapping method above. Thirdly, we align the chimeric-mapped FLNC with the protein-coding gene positions from the reference annotation, aggregate these alignment results, and thus determine the expression quantification (UMI counts) of protein-coding genes linked to each TE locus.

Supporting information

S1 Fig. Isoform switch caused functional/structural changes of genes. Expression pattern of gene isoforms which showed switch during preimplantation embryo development. The exact functional/structural domains predicted in each isoform are annotated based on the CATH database (http://cathdb.info/search/by_sequence).
(EPS)

S2 Fig. Relationship between gene and isoform expression. (A, B) The ratios of genes detected with different numbers of isoform types for each stage of mouse embryos and mESCs in this study (A) and SCAN-seq data (B). The raw data for these 2 plots are supplied in S6 Data. (C) The ratios of genes detected with different numbers of isoform types for full oocyte, 1/2 oocyte, and 1/4 oocyte. The raw data for this plot is supplied in S2 Data. (D, E) Expression levels of genes

detected with different numbers of isoform types for each stage of mouse embryos and mESCs in this study (D) and SCAN-seq data (E). The raw data for these 2 plots are supplied in S6 Data. (F) Gel view of cDNA amplification products of each gene. *Srsf7*, *Rps5*, and *Rps19* are examples of highly expressed genes (CPM >100) and *Plp2*, *Mrps6*, and *Ssl8l2* are lowly expressed genes (CPM <10). The raw image for this plot is supplied in S1 Raw Images. (G) Density plot showing the proportion of the major isoforms in genes expressing multiple isoform types. Only genes detected with UMI counts over 5 were included. The raw data for this plot is supplied in S6 Data.
(TIF)

S3 Fig. The characteristics of different types of transcripts. (A) Ratios of the transcripts overlapped with annotated TSS. Transcripts with the 5-terminal locating within 200 bp of the CAGE peaks are regarded as overlapped transcripts. The raw data for this plot is supplied in S7 Data. (B) Length distribution of different types of transcripts. The raw data for this plot is supplied in S7 Data. (C) Relative length of the predicted protein to the complete reference ORF of each type of transcript. The raw data for this plot is supplied in S7 Data. (D) Ratios of each type of transcript at different stages calculated using SCAN-seq data. The raw data for this plot is supplied in S7 Data. (E) Expression level of the 3-prime partial transcripts detected at each stage in SCAN-seq data. The raw data for this plot is supplied in S7 Data.
(EPS)

S4 Fig. Characteristics of TE expression during preimplantation embryo development. (A) Expression level of all TE counts belonging to different families in each stage. Each dot represents a cell. The raw data for this plot is supplied in S5 Data. Expression level of each TE locus (lower panel) belonging to different families in each stage. Only active loci in each cell were calculated. Each dot represents a cell. The raw data for this plot is supplied in S5 Data. (B) The top 5 subfamilies of ERVL expressed in clusters C4–C5 in Fig 5E. (C) The top 5 subfamilies of LINE1 expressed in clusters C4–C5 in Fig 5F.
(EPS)

S5 Fig. The loci expression of TEs and cell heterogeneity along developmental stages. (A) The loci expression of stage specific TEs (mean expression >1 log count across all cells) belonging to ERVL family. The raw data for this plot is supplied in S8 Data. (B) The loci expression of stage specific TEs (mean expression >1 log count across all cells) belonging to LINE1 family. The raw data for this plot is supplied in S8 Data. (C) Correlation coefficients of blastomeres within the same embryos or different embryos at the same stage base on gene expression data. The raw data for this plot is supplied in S8 Data. (D) Correlation coefficients of blastomeres within the same embryos or different embryos at the same stage base on isoform expression data. The raw data for this plot is supplied in S8 Data.
(EPS)

S1 Table. The list of stage-specific genes and stage-specific isoforms.
(XLSX)

S2 Table. The expression matrix of 3-prime partial transcripts.
(XLSX)

S3 Table. The expression matrix of TE loci.
(XLSX)

S4 Table. The list of TE chimeric genes.
(XLSX)

S1 Raw Images. Raw images.
(PDF)

S1 Data. Raw data for Fig 1.
(XLSX)

S2 Data. Raw data for Fig 2.
(XLSX)

S3 Data. Raw data for Fig 3.
(XLSX)

S4 Data. Raw data for Fig 4.
(XLSX)

S5 Data. Raw data for Figs 5 and S4.
(XLSX)

S6 Data. Raw data for S2 Fig.
(XLSX)

S7 Data. Raw data for S3 Fig.
(XLSX)

S8 Data. Raw data for S5 Fig.
(XLSX)

Acknowledgments

We thank Man Zhang (Guangzhou Laboratory) for teaching the methods to collect the preimplant embryos from mice. We thank Boyan Huang (Guangzhou Laboratory) for providing with the mESC cells. We thank Xining Li and Enze Deng (Guangzhou Laboratory) for helping analysis of some preliminary RNA-seq data.

Author Contributions

Conceptualization: Chuanle Xiao, Xiaoying Fan.

Data curation: Zhuoxing Shi, Xiaoying Fan.

Formal analysis: Zhuoxing Shi.

Funding acquisition: Xiaoying Fan.

Investigation: Chaoyang Wang, Qingpei Huang, Lei Chang, Xiaoying Fan.

Methodology: Zhuoxing Shi, Qingpei Huang, Rong Liu, Dan Su, Lei Chang.

Project administration: Chuanle Xiao, Xiaoying Fan.

Resources: Chaoyang Wang, Zhuoxing Shi, Rong Liu.

Software: Zhuoxing Shi.

Supervision: Chuanle Xiao, Xiaoying Fan.

Validation: Chaoyang Wang, Qingpei Huang, Dan Su.

Visualization: Zhuoxing Shi.

Writing – original draft: Chaoyang Wang, Zhuoxing Shi, Lei Chang, Xiaoying Fan.

Writing – review & editing: Chaoyang Wang, Zhuoxing Shi, Xiaoying Fan.

References

1. Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun. 2020; 11(1):4025. Epub 2020/08/14. https://doi.org/10.1038/s41467-020-17800-6 PMID: 32788667; PubMed Central PMCID: PMC7423900.
2. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021; 12(1):463. Epub 2021/01/21. https://doi.org/10.1038/s41467-020-20343-5 PMID: 33469025; PubMed Central PMCID: PMC7815907.
3. Oguchi Y, Ozaki Y, Abdelmoez MN, Shintaku H. NanoSINC-seq dissects the isoform diversity in subcellular compartments of single cells. Sci Adv. 2021; 7(15):eabe0317. https://doi.org/10.1126/sciadv.abe0317 PMID: 33827812
4. Fan X, Tang D, Liao Y, Li P, Zhang Y, Wang M, et al. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. PLoS Biol. 2020; 18(12):e3001017. Epub 2020/12/31. https://doi.org/10.1371/journal.pbio.3001017 PMID: 33378329; PubMed Central PMCID: PMC7773192.
5. Lebrigand K, Bergenstråhle J, Thrane K, Mollbrink A, Meletis K, Barbry P, et al. The spatial landscape of gene expression isoforms in tissue sections. Nucleic Acids Res. 2023; 51(8):e47–e. https://doi.org/10.1093/nar/gkad169 PMID: 36928528
6. Shi ZX, Chen ZC, Zhong JY, Hu KH, Zheng YF, Chen Y, et al. High-throughput and high-accuracy single-cell RNA isoform analysis using PacBio circular consensus sequencing. Nat Commun. 2023; 14(1):2631. Epub 2023/05/07. https://doi.org/10.1038/s41467-023-38324-9 PMID: 37149708; PubMed Central PMCID: PMC10164132.
7. Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-023-01815-7 PMID: 37291427
8. Liu Z, Roberts R, Mercer TR, Xu J, Sedlazeck FJ, Tong W. Towards accurate and reliable resolution of structural variants for clinical diagnosis. Genome Biol. 2022; 23(1):68. Epub 2022/03/05. https://doi.org/10.1186/s13059-022-02636-8 PMID: 35241127; PubMed Central PMCID: PMC8892125.
9. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics. 2016; 14(5):265–79. Epub 2016/10/31. https://doi.org/10.1016/j.gpb.2016.05.004 PMID: 27646134; PubMed Central PMCID: PMC5093776.
10. Xu Y, Zhao W, Olson SD, Prabhakara KS, Zhou X. Alternative splicing links histone modifications to stem cell fate decision. Genome Biol. 2018; 19(1):133. Epub 2018/09/16. https://doi.org/10.1186/s13059-018-1512-3 PMID: 30217220; PubMed Central PMCID: PMC6138936.
11. Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, et al. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. Proc Natl Acad Sci U S A. 2010; 107(23):10514–9. Epub 2010/05/26. https://doi.org/10.1073/pnas.0912260107 PMID: 20498046; PubMed Central PMCID: PMC2890851.
12. Rao S, Zhen S, Roumiantsev S, McDonald LT, Yuan GC, Orkin SH. Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. Mol Cell Biol. 2010; 30(22):5364–80. Epub 2010/09/15. https://doi.org/10.1128/MCB.00419-10 PMID: 20837710; PubMed Central PMCID: PMC2976381.
13. Jukam D, Shariati SAM, Skotheim JM. Zygotic Genome Activation in Vertebrates. Dev Cell. 2017; 42(4):316–32. Epub 2017/08/23. https://doi.org/10.1016/j.devcel.2017.07.026 PMID: 28829942; PubMed Central PMCID: PMC5714289.
14. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development. 2009; 136(18):3033–42. Epub 2009/08/25. https://doi.org/10.1242/dev.033183 PMID: 19700615.
15. Vastenhouw NL, Cao WX, Lipshitz HD. The maternal-to-zygotic transition revisited. Development. 2019; 146(11). Epub 2019/06/14. https://doi.org/10.1242/dev.161471 PMID: 31189646.
16. Zhao LW, Zhu YZ, Wu YW, Pi SB, Shen L, Fan HY. Nuclear poly(A) binding protein 1 (PABPN1) mediates zygotic genome activation-dependent maternal mRNA clearance during mouse early embryonic development. Nucleic Acids Res. 2022; 50(1):458–72. Epub 2021/12/15. https://doi.org/10.1093/nar/gkab1213 PMID: 34904664; PubMed Central PMCID: PMC8855302.
17. Qiao Y, Ren C, Huang S, Yuan J, Liu X, Fan J, et al. High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. Nat Commun. 2020; 11(1):2653. Epub 2020/05/29. https://doi.org/10.1038/s41467-020-16444-w PMID: 32461551; PubMed Central PMCID: PMC7253418.

18. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012; 487(7405):57–63. Epub 2012/06/23. https://doi.org/10.1038/nature11244 PMID: 22722858; PubMed Central PMCID: PMC3395470.

19. Yang M, Yu H, Yu X, Liang S, Hu Y, Luo Y, et al. Chemical-induced chromatin remodeling reprograms mouse ESCs to totipotent-like stem cells. Cell Stem Cell. 2022; 29(3):400–18 e13. Epub 2022/02/11. https://doi.org/10.1016/j.stem.2022.01.010 PMID: 35143761.

20. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002; 420(6915):520–62. Epub 2002/12/06. https://doi.org/10.1038/nature01262 PMID: 12466850.

21. Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, et al. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. Nat Struct Mol Biol. 2013; 20(3):332–8. Epub 2013/01/29. https://doi.org/10.1038/nsmb.2495 PMID: 23353788.

22. Berrens RV, Yang A, Laumer CE, Lun ATL, Bieberich F, Law CT, et al. Locus-specific expression of transposable elements in single cells with CELLO-seq. Nat Biotechnol. 2022; 40(4):546–54. Epub 2021/11/17. https://doi.org/10.1038/s41587-021-01093-1 PMID: 34782740.

23. Geis FK, Goff SP. Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview. Viruses. 2020; 12(8). Epub 2020/08/23. https://doi.org/10.3390/v12080884 PMID: 32823517; PubMed Central PMCID: PMC7472088.

24. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017; 18(2):71–86. https://doi.org/10.1038/nrg.2016.139 PMID: 27867194

25. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell. 2004; 7(4):597–606. Epub 2004/10/08. https://doi.org/10.1016/j.devcel.2004.09.004 PMID: 15469847.

26. Eckersley-Maslin MA, Svensson V, Krueger C, Stubbs TM, Giehr P, Krueger F, et al. MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell Rep. 2016; 17(1):179–92. Epub 2016/09/30. https://doi.org/10.1016/j.celrep.2016.08.087 PMID: 27681430; PubMed Central PMCID: PMC5055476.

27. Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. Cell. 2018; 174(2):391–405 e19. Epub 2018/06/26. https://doi.org/10.1016/j.cell.2018.05.043 PMID: 29937225; PubMed Central PMCID: PMC6046266.

28. Sun Z, Yu H, Zhao J, Tan T, Pan H, Zhu Y, et al. LIN28 coordinately promotes nucleolar/ribosomal functions and represses the 2C-like transcriptional program in pluripotent stem cells. Protein Cell. 2021. Epub 2021/08/01. https://doi.org/10.1007/s13238-021-00864-5 PMID: 34331666.

29. Yu H, Chen M, Hu Y, Ou S, Yu X, Liang S, et al. Dynamic reprogramming of H3K9me3 at hominoid-specific retrotransposons during human preimplantation development. Cell Stem Cell. 2022; 29(7):1031–50.e12. Epub 2022/07/09. https://doi.org/10.1016/j.stem.2022.06.006 PMID: 35803225.

30. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013; 10(11):1096–8. Epub 2013/09/24. https://doi.org/10.1038/nmeth.2639 PMID: 24056875.

31. Hardwick SA, Chen WY, Wong T, Deveson IW, Blackburn J, Andersen SB, et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat Methods. 2016; 13(9):792–8. Epub 2016/08/10. https://doi.org/10.1038/nmeth.3958 PMID: 27502218.

32. Paul L, Kubala P, Horner G, Ante M, Holländer I, Alexander S, et al. SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. bioRxiv. 2016:080747. https://doi.org/10.1101/080747

33. Molè MA, Weberling A, Zernicka-Goetz M. Comparative analysis of human and mouse development: From zygote to pre-gastrulation. Curr Top Dev Biol. 2020; 136:113–38. Epub 2020/01/22. https://doi.org/10.1016/bs.ctdb.2019.10.002 PMID: 31959285.

34. Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, et al. Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res. 2018; 46(D1):D1282. Epub 2017/12/02. https://doi.org/10.1093/nar/gkx1187 PMID: 29194501; PubMed Central PMCID: PMC5753360.

35. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Res. 2020; 49(D1):D266–D73. https://doi.org/10.1093/nar/gkaa1079 PMID: 33237325

36. Saha S, Matthews DA, Bessant C. High throughput discovery of protein variants using proteomics informed by transcriptomics. Nucleic Acids Res. 2018; 46(10):4893–902. Epub 2018/05/03. https://doi.org/10.1093/nar/gky295 PMID: 29718325; PubMed Central PMCID: PMC6007231.

37. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013; 500(7464):593–7. Epub 2013/07/31. https://doi.org/10.1038/nature12364 PMID: 23892778; PubMed Central PMCID: PMC4950944.

38. Yang Y, Wang L, Han X, Yang WL, Zhang M, Ma HL, et al. RNA 5-Methylcytosine Facilitates the Maternal-to-Zygotic Transition by Preventing Maternal mRNA Decay. Mol Cell. 2019; 75(6):1188–202 e11. Epub 2019/08/11. https://doi.org/10.1016/j.molcel.2019.06.033 PMID: 31399345.

39. Sha QQ, Zhu YZ, Li S, Jiang Y, Chen L, Sun XH, et al. Characterization of zygotic genome activation-dependent maternal mRNA clearance in mouse. Nucleic Acids Res. 2020; 48(2):879–94. Epub 2019/11/30. https://doi.org/10.1093/nar/gkz1111 PMID: 31777931; PubMed Central PMCID: PMC6954448.

40. Fu X, Zhang C, Zhang Y. Epigenetic regulation of mouse preimplantation embryo development. Curr Opin Genet Dev. 2020; 64:13–20. Epub 2020/06/22. https://doi.org/10.1016/j.gde.2020.05.015 PMID: 32563750; PubMed Central PMCID: PMC7641911.

41. Jia W, Yao Z, Zhao J, Guan Q, Gao L. New perspectives of physiological and pathological functions of nucleolin (NCL). Life Sci. 2017; 186:1–10. https://doi.org/10.1016/j.lfs.2017.07.025 PMID: 28751161

42. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020; 21(12):721–36. https://doi.org/10.1038/s41576-020-0251-y PMID: 32576954

43. Shahid S, Slotkin RK. The current revolution in transposable element biology enabled by long reads. Curr Opin Plant Biol. 2020; 54:49–56. Epub 2020/02/03. https://doi.org/10.1016/j.pbi.2019.12.012 PMID: 32007731.

44. Shen H, Yang M, Li S, Zhang J, Peng B, Wang C, et al. Mouse totipotent stem cells captured and maintained through spliceosomal repression. Cell. 2021; 184(11):2843–59 e20. Epub 2021/05/16. https://doi.org/10.1016/j.cell.2021.04.020 PMID: 33991488.

45. Song M, Pebworth MP, Yang X, Abnousi A, Fan C, Wen J, et al. Cell-type-specific 3D epigenomes in the developing human cortex. Nature. 2020; 587(7835):644–9. Epub 2020/10/16. https://doi.org/10.1038/s41586-020-2825-4 PMID: 33057195; PubMed Central PMCID: PMC7704572.

46. Mohanan NK, Shaji F, Koshre GR, Laishram RS. Alternative polyadenylation: An enigma of transcript length variation in health and disease. WIREs RNA. 2021; 13(1). https://doi.org/10.1002/wrna.1692 PMID: 34581021

47. Zhang W, Chen F, Chen R, Xie D, Yang J, Zhao X, et al. Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. Nucleic Acids Res. 2019; 47(16):8485–501. Epub 2019/07/16. https://doi.org/10.1093/nar/gkz594 PMID: 31304534; PubMed Central PMCID: PMC7145578.

48. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome Res. 2017; 27(8):1384–94. Epub 2017/05/20. https://doi.org/10.1101/gr.216150.116 PMID: 28522611; PubMed Central PMCID: PMC5538554.

49. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet. 2017; 49(10):1502–10. https://doi.org/10.1038/ng.3945 PMID: 28846101

50. Shigaki H, Baba Y, Watanabe M, Murata A, Iwagami S, Miyake K, et al. LINE-1 hypomethylation in gastric cancer, detected by bisulfite pyrosequencing, is associated with poor prognosis. Gastric Cancer. 2013; 16(4):480–7. Epub 2012/11/28. https://doi.org/10.1007/s10120-012-0209-7 PMID: 23179365; PubMed Central PMCID: PMC3824342.

51. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet. 2020; 52(3):306–19. Epub 2020/02/07. https://doi.org/10.1038/s41588-019-0562-0 PMID: 32024998; PubMed Central PMCID: PMC7058536.

52. Zhang R, Zhang F, Sun Z, Liu P, Zhang X, Ye Y, et al. LINE-1 Retrotransposition Promotes the Development and Progression of Lung Squamous Cell Carcinoma by Disrupting the Tumor-Suppressor Gene FGGY. Cancer Res. 2019; 79(17):4453–65. Epub 2019/07/11. https://doi.org/10.1158/0008-5472.CAN-19-0076 PMID: 31289132.

53. Liu X, Liu Z, Wu Z, Ren J, Fan Y, Sun L, et al. Resurrection of endogenous retroviruses during aging reinforces senescence. Cell. 2023; 186(2):287–304 e26. Epub 2023/01/08. https://doi.org/10.1016/j.cell.2022.12.017 PMID: 36610399.

54. Modzelewski AJ, Shao W, Chen J, Lee A, Qi X, Noon M, et al. A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform essential for development. Cell. 2021; 184(22):5541–58.e22. Epub 2021/10/14. https://doi.org/10.1016/j.cell.2021.09.021 PMID: 34644528; PubMed Central PMCID: PMC8787082.

55. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol. 2022; 23(7):481–97. Epub 2022/03/02. https://doi.org/10.1038/s41580-022-00457-y PMID: 35228718; PubMed Central PMCID: PMC10470143.

56. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014; 516(7531):405–9. Epub 2014/10/16. https://doi.org/10.1038/nature13804 PMID: 25317556.

57. Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell. 2015; 161(6):1453–67. Epub 2015/06/06. https://doi.org/10.1016/j.cell.2015.04.053 PMID: 26046444; PubMed Central PMCID: PMC4459712.

58. Du Z, Zheng H, Huang B, Ma R, Wu J, Zhang X, et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. Nature. 2017; 547(7662):232–5. Epub 2017/07/14. https://doi.org/10.1038/nature23263 PMID: 28703188.