Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data

Wanqing Shao[1,2] and Ting Wang[1,2,3]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA; [2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri 63110, USA; [3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri 63108, USA

Transposable elements (TEs) are an integral part of the host transcriptome. TE-containing noncoding RNAs (ncRNAs) show considerable tissue specificity and play important roles during development, including stem cell maintenance and cell differentiation. Recent advances in single-cell RNA-seq (scRNA-seq) revolutionized cell type–specific gene expression analysis. However, effective scRNA-seq quantification tools tailored for TEs are lacking, limiting our ability to dissect TE expression dynamics at single-cell resolution. To address this issue, we established a TE expression quantification pipeline that is compatible with scRNA-seq data generated across multiple technology platforms. We constructed TE-containing ncRNA references using bulk RNA-seq data and showed that quantifying TE expression at the transcript level effectively reduces noise. As proof of principle, we applied this strategy to mouse embryonic stem cells and successfully captured the expression profile of endogenous retroviruses in single cells. We further expanded our analysis to scRNA-seq data from early stages of mouse embryogenesis. Our results illustrated the dynamic TE expression at preimplantation stages and revealed 146 TE-containing ncRNA transcripts with substantial tissue specificity during gastrulation and early organogenesis.

[Supplemental material is available for this article.]

Transposable elements (TEs) occupy a large proportion of eukaryotic genomes, representing ~50% of the human genome and 40% of the mouse genome (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002). Although once regarded as nonfunctional parasitic sequences, increasing evidence suggests that TE-derived sequences play pivotal roles in gene regulation. During evolution, TEs rewire host transcription networks through transposition and co-option, resulting in a wide variety of TE-derived regulatory elements, including promoters, enhancers, transcription terminators, and chromatin loop anchors (for reviews, see Feschotte and Gilbert 2012; Rebollo et al. 2012; Cowley and Oakey 2013; Garcia-Perez et al. 2016; Chuong et al. 2017; Sundaram and Wysocka 2020). In the present day, despite losing most of their transposition abilities, TE-derived sequences continue to impact host genomes through transcription, which generates protein-coding TE chimeric RNAs as well as noncoding RNAs (ncRNAs) that are involved in normal and cancer development (for reviews, see Gifford et al. 2013; Hadjiargyrou and Delihas 2013; Hutchins and Pei 2015; Anwar et al. 2017; Rodriguez-Terrones and Torres-Padilla 2018).

TEs are major contributors of ncRNAs in both human and mouse. More than two-thirds of mature long ncRNAs contain at least one TE and almost half of the total base pairs of long ncRNA are derived from TEs (Kelley and Rinn 2012; Kapusta et al. 2013). TE-containing ncRNAs show substantial developmental stage and tissue specificity and participate in embryonic stem cell (ESC) maintenance and early embryogenesis. For instance, endogenous retroviruses (ERVs) are highly expressed in ESCs and ERV-derived transcripts are involved in the maintenance of pluripotency (Macfarlan et al. 2012; Santoni et al. 2012; Fort et al. 2014; Lu et al. 2014; Ohnuki et al. 2014; Wang et al. 2014). During mouse and human embryogenesis, a large number of TEs, including ERVs, long interspersed nuclear element-1 (LINE-1), and short interspersed elements (SINEs) become active and contribute to a significant proportion of total RNAs before blastocyst stage (Kigami et al. 2003; Peaston et al. 2004; Svoboda et al. 2004; Maksakova and Mager 2005; Fadloun et al. 2013; Göke et al. 2015; Grow et al. 2015; De Iaco et al. 2017; Ge 2017; Hendrickson et al. 2017; Jachowicz et al. 2017; Whiddon et al. 2017; Percharde et al. 2018). Moreover, knocking down specific TE families, including LINE-1 and MuERV-L, results in clear developmental defects (Kigami et al. 2003; Huang et al. 2017; Jachowicz et al. 2017; Percharde et al. 2018).

Despite the importance of TEs, quantifying TE expression using high-throughput sequencing data has been challenging. Owing to the repetitive nature of TEs, sequencing reads that overlap with TEs are often discarded as a result of ambiguous mapping. Several software tools have been developed to address this issue, and they enabled TE expression quantification in bulk RNA-seq data (Criscione et al. 2014; Jin et al. 2015; Lerat et al. 2017; Jeong et al. 2018; Bendall et al. 2019; Kong et al. 2019; Yang et al. 2019). To quantify the expression of repetitive elements, these tools often aggregate multialigned reads at TE subfamilies/families or redistribute them to individual TEs based on heuristic or statistical rules. Although proven to be successful in a range of

Corresponding author: twang@wustl.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.265173.120.

31:88–100 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/21; www.genome.org

biological systems, applications of the current TE quantification strategies were mostly limited to bulk RNA-seq, which lacks the ability to distinguish cell type–specific TE expression.

Recent developments in single-cell RNA-seq (scRNA-seq) provide unprecedented opportunities for examining cell type–specific TE expression. However, effective TE quantification tools optimized for scRNA-seq data are lacking. Although the assessment of genome-wide transcriptional activity of TEs in single cells has been attempted by counting signals at individual TE fragments or TE subfamilies/families (Göke et al. 2015; Ge 2017; Boroviak et al. 2018; Brocks et al. 2018; Yandım and Karakülah 2019; He et al. 2020; Jonsson et al. 2020), such approaches are not optimal. Compared with bulk RNA-seq, scRNA-seq signal is much noisier and often shows 5′ or 3′ end enrichment along the transcripts. Counting reads at individual TEs or subfamilies/families fails to take into account the structures of the full-length transcripts, which can consist of multiple TEs from different subfamilies/families. Consequently, different expression values will be assigned to individual TEs within the same transcript. This caveat is especially obvious when dealing with scRNA-seq data sets in which sequencing reads are enriched at either the 5′ or 3′ end of the RNA. Counting reads without the knowledge of the full-length transcripts will only capture TEs near the 5′ end or the poly(A) signal, resulting in an inaccurate picture of the genome-wide TE expression pattern (O'Neill et al. 2020).

In this work, we present an analytical framework tailored to TE expression quantification in scRNA-seq data sets. We systematically evaluated scRNA-seq reads that mapped to TEs and showed that quantifying TE expression in single cells using transcripts assembled from bulk RNA-seq effectively reduces noise. Applying our strategy to mouse early embryogenesis illustrated the dynamic TE expression during preimplantation stages and revealed TE-containing ncRNAs with substantial tissue enrichment during gastrulation and early organogenesis.

Results

A higher percentage of reads are mapped to TEs in scRNA-seq compared with bulk RNA-seq

Owing to the biological significance of TE-containing ncRNAs, we decided to focus our analysis on TEs that are not part of the exons of protein-coding genes. First, to determine the fraction of reads that can be mapped to these TEs in scRNA-seq data, we processed 36 publicly available single-cell data sets (Supplemental Table S1). These data sets contain both human and mouse samples and were generated using seven different scRNA-seq protocols. Bulk RNA-seq data sets from the same study or derived from the same cell line were included as controls. To preserve reads that originate from repetitive regions, multiple mapping was implemented during the alignment step. Reads that mapped to multiple locations or overlapped with more than one feature were distributed equally for signal quantification. Calculating the number of mappable reads based on genomic locations revealed that a large proportion of reads overlap with TEs in all tested scRNA-seq data sets, suggesting that TE expression can be captured by scRNA-seq (Fig. 1A). We also observed that a higher percentage of reads was mapped to TEs in scRNA-seq compared with bulk RNA-seq. This phenomenon was consistent across different scRNA-seq platforms even when only uniquely mapped reads were considered (Fig. 1A; Supplemental Fig. S1A), suggesting that the high TE mapping ratio is prevalent in scRNA-seq data sets and not introduced by non-unique mapping.

To evaluate whether the high TE mapping percentage in scRNA-seq was associated with data quality per cell, we further compared scRNA-seq data generated using Smart-seq, 10x Genomics Chromium, SCRB-seq, and Drop-seq and examined the relationships between the percentage of TE reads per cell and the two key parameters indicative of scRNA-seq quality: sequencing depth and the percentage of mitochondrial reads (Ilicic et al. 2016). Our analysis revealed that a high TE mapping percentage was observed across individual cells with limited correlation to sequencing depth (Supplemental Fig. S1B,C). Similarly, no strong correlation was detected between the percentage of TE reads and that of the mitochondrial reads (Supplemental Fig. S1C). These results suggest that the high TE mapping ratio in scRNA-seq is unlikely to be an artifact caused by variations in sequencing depth or cell death.

To address the concern that genomic DNA contamination may contribute to the majority of TE signal in scRNA-seq, we next quantified the number of total reads mapped to five nonoverlapping genomic regions: protein-coding exons, TEs within the introns of protein-coding genes, other intronic regions of protein-coding genes, intergenic TEs, and other intergenic regions. A higher percentage of total reads was mapped to the intronic regions of protein-coding genes in scRNA-seq, and the majority of TE overlapping reads were located within introns (Supplemental Fig. S1A), arguing against severe genomic DNA contamination.

Although additional experiments and analyses will be needed to pinpoint the origin of these intronic reads in scRNA-seq data, considering their presence across diverse scRNA-seq platforms, the limited correlation between the amount of intronic reads and sequencing quality, as well as previous successes in using intronic reads to infer transcription dynamics and cell states (Gaidatzis et al. 2015; La Manno et al. 2018), we hypothesize that these intronic reads originate from unprocessed RNA. Indeed, consistent with previous reports (La Manno et al. 2018; Selewa et al. 2020), we observed regions that are enriched for unspliced scRNA-seq reads and located within introns that tend to be flanked by AT-rich sequences, which could be involved in the poly(A) priming step during cDNA synthesis (Supplemental Fig. S1D).

Counting scRNA-seq reads at individual TEs leads to large numbers of false positive candidates

Current TE expression analyses often quantify RNA-seq signal at individual TE fragments or TE subfamilies/families (Criscione et al. 2014; Jin et al. 2015; Lerat et al. 2017; Jeong et al. 2018; Yang et al. 2019; He et al. 2020; Jonsson et al. 2020). Our observation that a large proportion of scRNA-seq reads map to TEs, especially intronic TEs, raises the concern that counting reads at single TEs or TE subfamilies/families will aggregate noise and fail to exclude TEs that are part of protein-coding genes, resulting in high numbers of false positive candidates. To test this, we applied a similar strategy and analyzed bulk and Smart-seq data sets generated using mouse embryonic stem cells (mESCs) cultured in 2i medium (Buettner et al. 2015; Kolodziejczyk et al. 2015). Because mESCs represent a population with limited heterogeneity and reads generated by Smart-seq and bulk RNA-seq share a similar distribution along the gene body (Ramsköld et al. 2012), we expected that the expression profiles obtained with scRNA-seq to be largely similar to those generated with bulk RNA-seq.

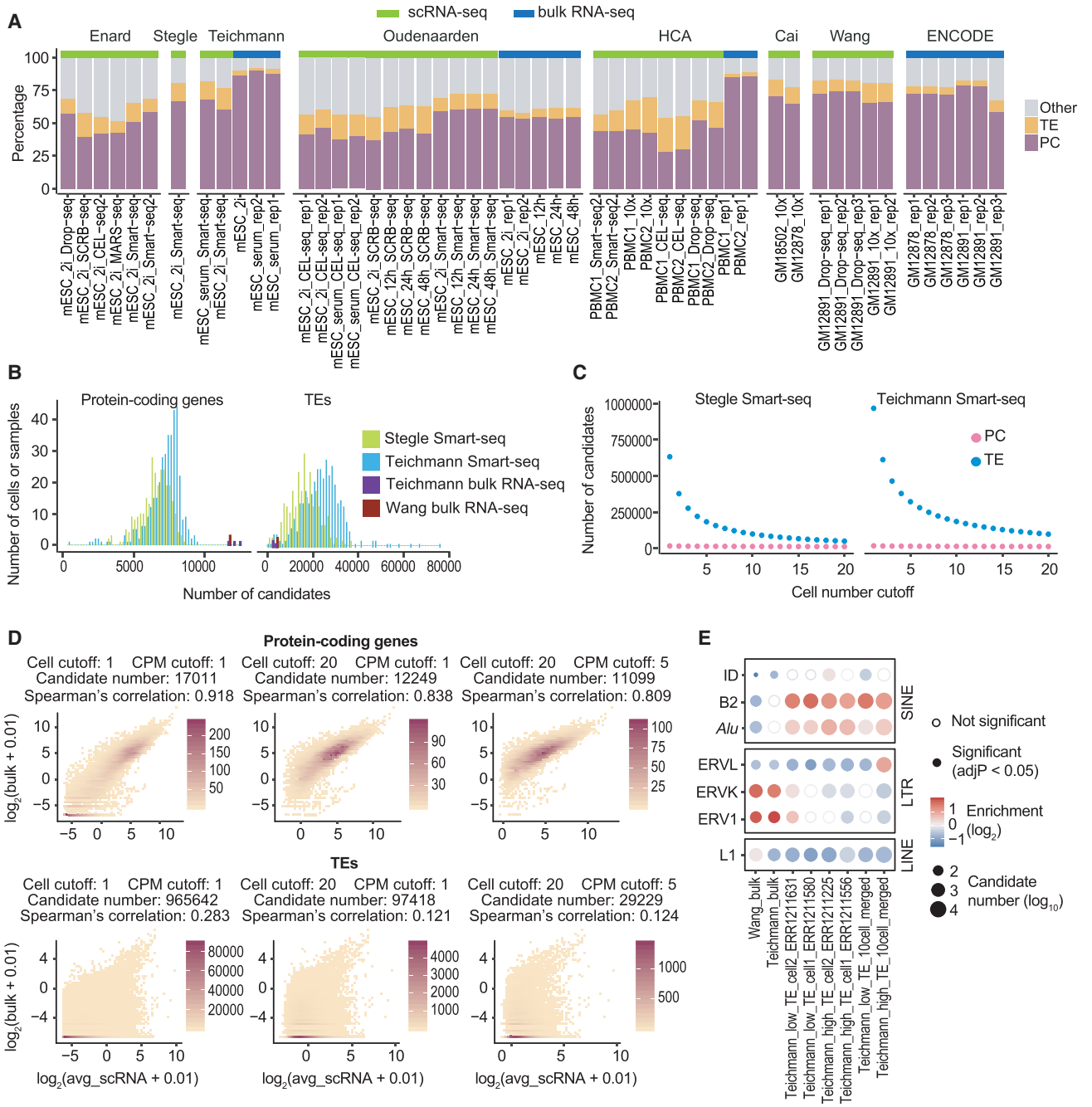

Figure 1. Counting scRNA-seq signal at individual TEs results in large numbers of false positive candidates. (*A*) Distribution of mappable reads in 16 bulk RNA-seq and 36 scRNA-seq data sets. Compared to bulk RNA-seq, scRNA-seq data have a higher percentage of reads mapped to TEs. Samples were arranged by studies. Data sets used in this figure are summarized in Supplemental Table S1. (PC) Protein-coding exons defined by RefSeq; (TE) transposable elements that do not overlap with protein-coding exons; (Other) other genomic locations; (mESC) mouse embryonic stem cell; (PBMC) human peripheral blood mononuclear cell; (GM12878 and GM12891) human lymphoblastoid cell lines. (*B*) Number of expressed (counts per million, CPM ≥ 1) protein-coding genes and TEs in mESC bulk RNA-seq and Smart-seq samples. On average, 12,000 protein-coding genes and 6000 TEs were detected in each bulk RNA-seq sample. In contrast, scRNA-seq captured 7000 protein-coding genes and 20,000 TEs per cell. (*C*) Number of candidates as a function of cell number cutoff. (Cell number cutoff) Minimum number of cells each candidate is expressed in; (expression cutoff) CPM ≥ 1. A cell number cutoff of 10 requires a candidate to have at least 1 CPM in at least 10 cells. Although the majority of protein-coding gene candidates were consistently detected in mESC Smart-seq data, a large number of TE candidates were detected in fewer than 10 cells. (*D*) Correlation between bulk RNA-seq and averaged scRNA-seq signal at protein-coding genes and TEs (Teichmann laboratory, mESC). Low correlation between bulk RNA-seq and averaged Smart-seq signal was observed at TEs regardless of expression cutoff. (Cell cutoff) Minimum number of cells each candidate is expressed in; (CPM cutoff) minimum CPM value for one candidate to be considered as expressed. Color scale represents the number of candidates. (*E*) TE-family enrichment analysis using TE candidates identified from mESC bulk RNA-seq and Smart-seq. Enrichment of ERV elements was observed with bulk RNA-seq data, but not in single cells. Smart-seq data of four single cells with different percentage of TE reads and merged Smart-seq data from 10 cells were included.

We first calculated the numbers of expressed protein-coding genes and expressed TEs (CPM ≥ 1) in these data sets. On average, 12,000 protein-coding genes and 6000 TEs were detected in bulk RNA-seq samples. In contrast, scRNA-seq captured an average of 7000 protein-coding genes and 20,000 TEs per cell (Fig. 1B). To evaluate the quality of these expressed candidates, we examined the following three parameters: (1) the number of cells each candidate is expressed in, (2) the correlation between the signal in bulk RNA-seq and the average signal across single cells, and (3) for TE candidates, the overrepresented TE families among all candidates. We reasoned that a candidate representative to the population should be expressed in a relatively large number of mESCs and show a strong correlation between its bulk RNA-seq and averaged scRNA-seq signal. However, only protein-coding genes matched this expectation (Fig. 1C,D; Supplemental Fig. S2A). A large proportion of TE candidates were only detected in a small number of cells and showed weak correlations between scRNA-seq and bulk RNA-seq signal regardless of the expression cutoff. This observation remained valid after we performed the same analysis by counting signals from individual exons. The exon length distributions were comparable to those of TEs, ruling out the possibility that length discrepancy between TEs and protein-coding genes contributes to false positive TE candidates (Supplemental Fig. S2B–D). We further compared overrepresented TEs within candidates identified from bulk RNA-seq and scRNA-seq by performing a TE-family enrichment analysis (Supplemental Fig. S3A). Although ERV1 and ERVK elements have been shown to be expressed in stem cells (Santoni et al. 2012; Fort et al. 2014; Lu et al. 2014; Ohnuki et al. 2014; Wang et al. 2014), they were only enriched in bulk RNA-seq in this analysis (Fig. 1E). scRNA-seq candidates obtained from this analysis were depleted of ERV1 and ERVK and instead enriched for SINEs (Fig. 1E; Supplemental Fig. S3B), which are often found near protein-coding genes and provide sequences that could act as reverse transcription priming sites (Medstrand et al. 2002).

In summary, the high number of TE candidates obtained from scRNA-seq, the weak signal correlation between individual cells, as well as the discordance between bulk and scRNA-seq strongly suggest that counting scRNA-seq reads at individual TEs will result in large numbers of false positive candidates.

Transcript assembly improves TE expression analysis

Transcript annotation serves as the cornerstone for expression quantification. Our ability to accurately assess the expression of protein-coding genes relies on well-annotated gene structures, which help to focus analysis on genomics regions with true signal. Although individual TEs are well annotated, it is usually unclear which TEs are expressed in a biological system and what the underlying transcript structures are. We reason that the large number of false positive candidates in scRNA-seq analysis is caused by counting sparse and noisy signal at millions of TE copies, of which only a small proportion are truly expressed (Supplemental Fig. S3C). Indeed, the signal correlation between averaged scRNA-seq and bulk RNA-seq is much stronger at TEs that overlap with the exons of RefSeq annotated ncRNA (Supplemental Fig. S4A). Therefore, we hypothesize that incorporating transcript structures into the analysis should help to reduce noise.

Several recent studies took advantage of well-studied TE transcription units such as full-length ERVs or LINEs for TE expression quantification, but did not consider transcripts that are composed of TEs from different families or classes (Tokuyama et al. 2018; Bendall et al. 2019; McKerrow and Fenyö 2020). To obtain a more comprehensive catalog of ncRNAs with exonized TEs, we performed transcript assembly using mESC bulk RNA-seq data. We selected transcripts with lengths exceeding 200 nt and identified 692 transcripts whose exons overlap with TEs but not the exons of RefSeq annotated protein-coding genes (Fig. 2A,B; Supplemental Fig. S5). These include transcripts that are entirely derived from TEs (e.g., some ERV transcripts and LINE transcripts) as well as transcripts derived from multiple fragmented TEs and TE–non-TE hybrid units. These transcripts were termed TE transcripts. To test the accuracy of our assembly, we focused on the promoters of assembled TE transcripts and examined several genomic signatures that are indicative of active transcription. Indeed, the majority of our TE transcript promoters overlap with FANTOM5 CAGE peaks (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014) and are enriched for ATAC-seq signal while depleted of CpG methylation (Fig. 2C).

Using these newly generated transcript models, we recalculated the expression of the TE transcripts. Because a major source of noise in measuring TE expression comes from intronic reads (Supplemental Fig. S4A) and intronic signals are products of passive cotranscription of mature RNAs (Lanciano and Cristofari 2020), we only quantified signals within the exonic regions of these TE transcripts. Quantifying TE expression at the TE transcripts led to a much stronger correlation between mESC bulk and Smart-seq data (Fig. 2D; Supplemental Fig. S4B). Furthermore, we obtained much more consistent TE-family enrichment results between bulk and scRNA-seq and were able to identify the expression of ERV elements at single cells (Fig. 2E,F).

Although Smart-seq-based protocols generate deeper sequencing depth with reads covering the entire transcript, other popular scRNA-seq strategies such as 10x Genomics Chromium, Drop-seq, and SCRB-seq produce shallow sequencing with reads biased toward the 5′ or 3′ end of the RNA (Ramsköld et al. 2012; Soumillon et al. 2014; Macosko et al. 2015; Zheng et al. 2017). Counting reads at individual TEs using data with 5′ or 3′ signal enrichment will only capture TEs that are located at either end of the transcripts, thus biasing our understanding about TE expression. We reason that our approach can help to overcome this limitation by using the annotation of full-length TE transcripts.

To support our reasoning, we analyzed the data set from a previously published mESC differentiation study, in which single-cell Smart-seq2, single-cell SCRB-seq, and bulk RNA processed with SCRB-seq protocol were performed (Semrau et al. 2017). Using individual TEs as reference, we observed a severe discordance of TE expression between SCRB-seq and Smart-seq2, likely resulting from differences in signal distribution along the transcripts (Supplemental Fig. S6A). Conversely, using full-length TE transcripts as reference led to significantly improved signal correlations (Supplemental Fig. S6B). Moreover, quantifying TE expression at transcript level allowed us to recover the enrichment of ERVs from all data sets, whereas only Smart-seq2 showed ERV enrichments when counting at individual TEs (Supplemental Fig. S6C).

Taken together, these results suggest that our analysis approach is applicable not only to scRNA-seq data with a high number of reads covering the entire transcript body, but also to other popular scRNA-seq strategies that feature shallow sequencing depth at the 3′ end of the transcripts.

Finally, we examined whether the incorporation of an expectation–maximization (EM) algorithm will improve the accuracy of expression quantification at repetitive regions. We simulated RNA-seq reads at TE transcripts with a wide range of coverages and

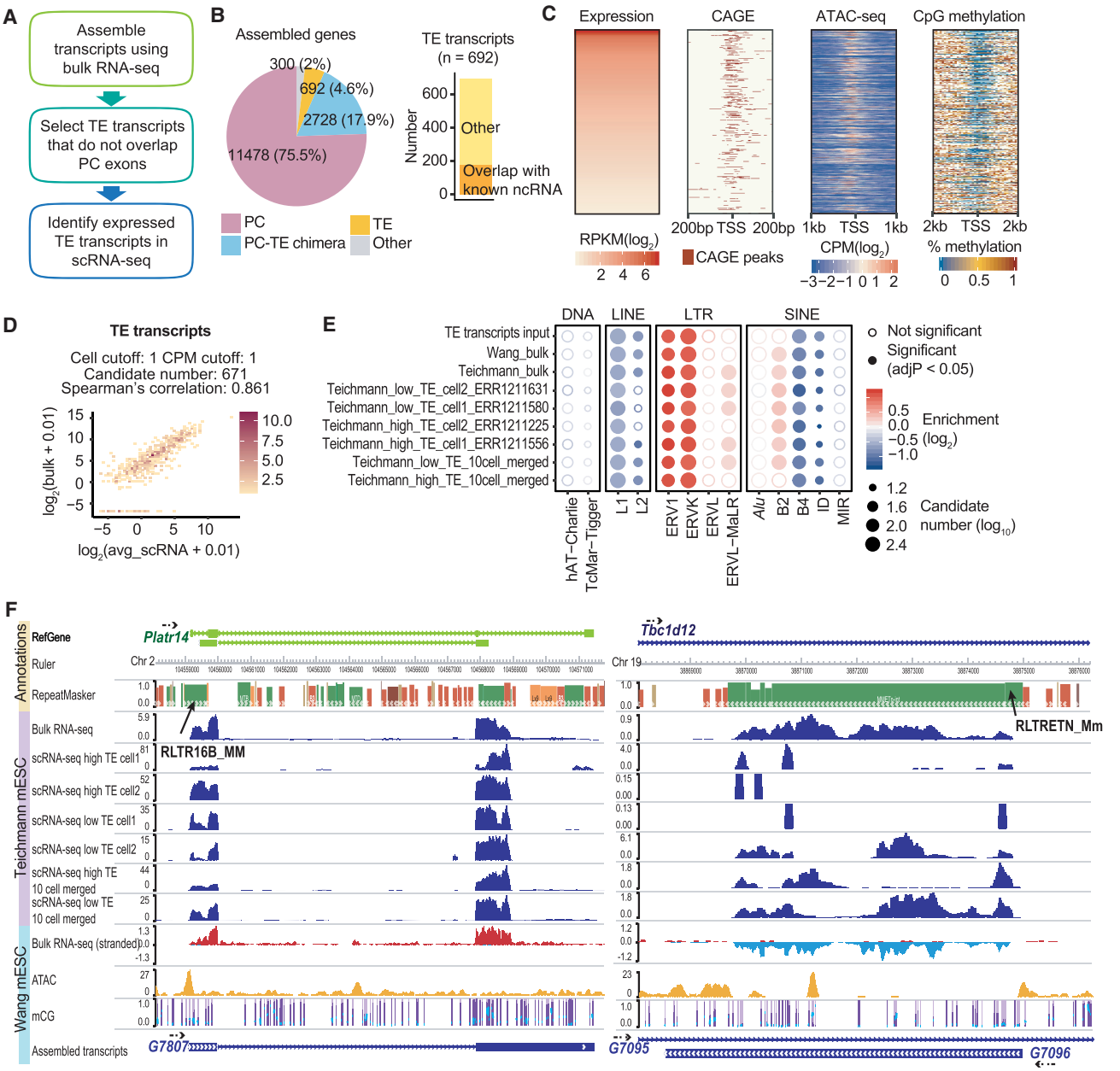

Figure 2. Transcript assembly improves scRNA-seq TE expression analysis. Data sets used in this figure are summarized in Supplemental Table S1. (*A*) Flowchart of scRNA-seq TE quantification pipeline. In short, transcript assembly was performed with bulk RNA-seq data, and transcripts that overlap with TEs but not protein-coding exons were used for expression quantification in scRNA-seq data. (*B*) Transcript assembly using three mESC bulk RNA-seq data (Wang laboratory) yielded 692 TE transcripts. Among these TE transcripts, 179 overlap with ncRNAs annotated by RefSeq. (*C*) FANTOM5 CAGE peaks, ATAC-seq signals, and CpG methylation signals at the promoter region of TE transcripts with RPKM ≥ 1 (reads per kilobase million). (*D*) Correlation between mESC bulk RNA-seq and averaged Smart-seq (Teichmann laboratory) signals at TE transcripts. Color scale represents the number of candidates. (*E*) TE-family enrichment analysis using expressed TE transcripts. Enrichment of ERV elements was observed with both bulk RNA-seq and Smart-seq samples. (*F*) Examples of TE transcript. Assembled TE transcripts, uniquely mapped reads of mESC bulk RNA-seq, Smart-seq, merged Smart-seq, ATAC-seq, and CpG methylation were included. (*Left*) A TE transcript that initiates from RLTR16b_MM. This TE transcript overlaps *Platr14*, a long ncRNA known to impact the mESC differentiation-associated genes. (*Right*) A TE transcript that initiates from RLTRETN_Mm. This transcript is largely composed of TEs and reflect the transcription unit of ERV.

quantified the observed signal by redistributing multiple mapped reads with an EM algorithm using the amount of uniquely mapped reads as priors. In line with previous reports (Jin et al. 2015; Yang et al. 2019), we found that redistributing multiple mapped reads using the EM algorithm outperforms even-distribution, resulting in a higher percentage of observed reads matching the ground truth (Supplemental Fig. S7). Based on these observations, we implemented this EM-based algorithm in the expression quantification step.

Dynamic TE expression in preimplantation embryos

Encouraged by the results from the mESC data, we decided to apply our strategy to a more complex biological system: mouse

embryogenesis. The dynamic regulation of the epigenome during development not only fine-tunes protein-coding genes, but also allows specific TE expression at different developmental stages (Rowe and Trono 2011; Gifford et al. 2013; Gerdes et al. 2016; Rodriguez-Terrones and Torres-Padilla 2018; Deniz et al. 2019). Several recent studies used scRNA-seq to profile the transcription landscape of mouse embryos from zygote to early organogenesis, providing valuable resources for dissecting the dynamic expression of TEs.

To facilitate TE expression quantification, we performed transcript assembly using 37 bulk RNA-seq samples (Supplemental Table S1) that cover a range of tissues and developmental stages and obtained 5299 TE transcripts (Fig. 3A,B; Supplemental Fig. S8A; Supplemental Table S2). Seven hundred seventy of these transcripts overlap with known ncRNAs annotated by RefSeq (Fig. 3A). Compared with assembled protein-coding transcripts, which show similar length and exon number as RefSeq protein-coding gene annotations, assembled TE transcripts are shorter in length and possess fewer exons, a pattern consistent with annotated ncRNAs (Supplemental Fig. S8B,C).

Next, we analyzed three publicly available data sets in which the transcription landscape of mouse embryos from zygote to gastrulation was profiled using Smart-seq derived protocols (Deng et al. 2014; Mohammed et al. 2017; Cheng et al. 2019). In these data sets, a significant number of reads overlap with TEs, and ~3% of the total reads are mapped to TE transcripts (Supplemental Fig. S9A,B). After data integration and dimension reduction using the top 4000 variable features, we observed clear clustering patterns that were driven by cell type and developmental stage (Fig. 3C). Among the top 4000 variable features, 410 are TE transcripts, suggesting that the expression of TE transcripts could be cell type– or developmental stage–specific (Supplemental Fig. S9C–E). Indeed, we were able to observe TE transcript expression with strong tissue or stage specificity (Fig. 3D).

To further investigate the dynamics of TE transcription, we focused on preimplantation stages, in which high TE expression was documented. Because of the limited number of cells, scRNA-seq signals of each TE transcript across all the cells with the same developmental stage were averaged to reduce noise. Grouping TE transcripts based on their expression patterns across preimplantation stages resulted in the following six clusters (Fig. 3E): TE transcripts that are maternally deposited (cluster 1), TE transcripts that are up-regulated during minor and major waves of zygotic genome activation (clusters 2 and 3), TE transcripts that are up-regulated during zygotic genome activation and keep accumulating until the blastocyst stage (cluster 4), and TE transcripts that are up-regulated in the early- and mid-blastocyst stage (clusters 5 and 6).

We next performed TE enrichment analysis and observed that TE transcripts with distinct expression profiles tend to be enriched for different TE subfamilies (Fig. 3F). For instance, IAP elements are highly enriched in cluster 4, consistent with a previous report that IAP expression initiates from the two-cell stage, accumulates, and then disappears at the blastocyst stage (Pikó et al. 1984; Poznanski and Calarco 1991; Svoboda et al. 2004). We also observed the enrichment of ERVL and ERVL-MaLR members in cluster 2, in line with previous studies suggesting that ERVL and ERVL-MaLR members are highly expressed during the two-cell stage and constitute ~5% of the total transcripts (Kigami et al. 2003; Peaston et al. 2004; Svoboda et al. 2004). MTA_Mm-int and ORR1B1-int from the ERVL-MaLR family were also enriched in cluster 6, showing high expression during E4 blastocyst stage, an intriguing observation that is yet to be validated. Moreover, transcription factor binding site analysis using a 500-bp window upstream of TE transcripts identified footprints of transcription factors that were shown to be involved in mouse early embryogenesis such as Krupple-like factors, GABPA, and ELF5 (Ristevski et al. 2004; Donnison et al. 2005; Zhou et al. 2005; Bialkowska et al. 2017), suggesting shared regulatory networks between TE transcripts and protein-coding genes (Supplemental Fig. S10).

Tissue-specific TE expression during mouse gastrulation and early organogenesis

Comparing to preimplantation stages, TE expression during gastrulation and organogenesis is much less well studied, and a comprehensive catalog of tissue-specific TE transcripts is lacking. To address this, we analyzed a 10x scRNA-seq data set in which more than 100,000 cells were assayed using mouse E6.5 to E8.5 embryos (Pijuan-Sala et al. 2019). Comparing with mESC or mouse preimplantation data analyzed in previous sections, this E6.5 to E8.5 10x data contains considerably fewer TE overlapping reads, with ~1% of the UMI mapping to TE transcripts (Supplemental Fig. S11A,B). Although TE transcripts are in general lowly expressed and lack the high standardized variance observed at some protein-coding genes, they still constitute a small proportion of the top 1000 variable features that can be used to recapitulate the clustering pattern in the original study (Fig. 4A,B; Supplemental Fig. S11C). Furthermore, we were able to observe TE expression patterns that are enriched in small clusters of cells, suggesting that TE transcripts display considerable tissue specificity during these stages (Fig. 4C).

Next, we systematically examined the dynamic TE expression and obtained 146 TE transcripts that show substantial tissue enrichment (Supplemental Table S3). Hierarchical clustering analysis using the expression of these TE transcripts showed that tissues with similar origins are grouped together (Fig. 4D). For instance, tissues within the hematoendothelial lineage including hematoendothelial progenitors, endothelium, blood progenitors, and erythroids are adjacent to each other, and tissues linked to the neuronal lineage including neuromesodermal progenitor, spinal cord, forebrain/midbrain/hindbrain, and neural crest are clustered together.

Although 10x reads were enriched at the 3′ end of the transcript, all TEs located along the transcripts were captured using our assembled full-length TE transcripts (Fig. 4E). Among the 146 TE transcripts, 84 initiated from TE or have >50% of their exonic sequences contributed by TEs (Fig. 4D; Supplemental Fig. S11D). Overlapping these 146 TE transcripts with annotated ncRNAs revealed that 51 have been annotated by RefSeq. Although these known ncRNAs tend to contain a lower percentage of TEs, transcripts that are almost exclusively composed of TEs are largely unannotated, demonstrating the value of our approach in capturing transcripts that originated from highly repetitive regions. Moreover, we observed that TE transcripts enriched in different tissues display distinct sequence composition. For instance, TE transcripts enriched in extraembryonic ectoderm, extraembryonic endoderm, parietal endoderm, and visceral endoderm are almost exclusively composed of LTRs, although we did find that several highly expressed transcripts in extraembryonic ectoderm are derived from LINEs (Supplemental Fig. S12).

Discussion

Current genome-wide TE expression quantification tools often count signal at individual TEs or TE subfamilies/families. This

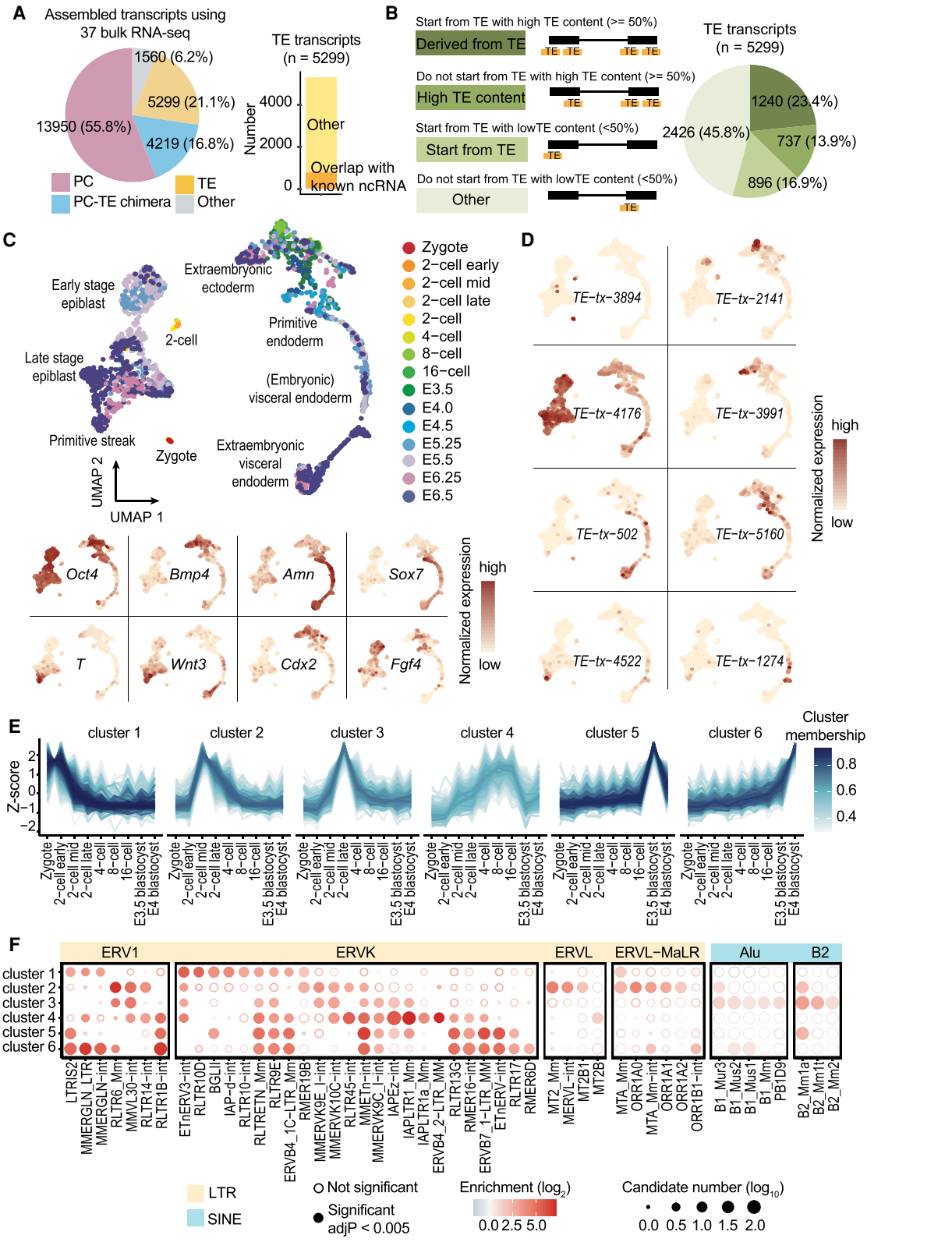

Figure 3. Dynamic TE expression in mouse preimplantation embryos. (*A*) Using 37 bulk RNA-seq samples, 5299 TE transcripts were constructed. Of these, 770 TE transcripts overlap with ncRNAs annotated by RefSeq. (*B*) More than half of all the assembled TE transcripts either initiate from TEs or have >50% of their exons composed of TEs. (*C*, *upper*) UMAP of scRNA-seq data from mouse zygote to E6.5 embryos. Cells were colored based on developmental stages. (*Lower*) Expression of cell type–specific markers. (*D*) Examples of developmental stage– and tissue-specific TE transcripts. (*E*) TE transcripts were grouped into six clusters based on their expression pattern across preimplantation stages. (*F*) TE subfamily enrichment analysis using TE transcripts within each of the six clusters.

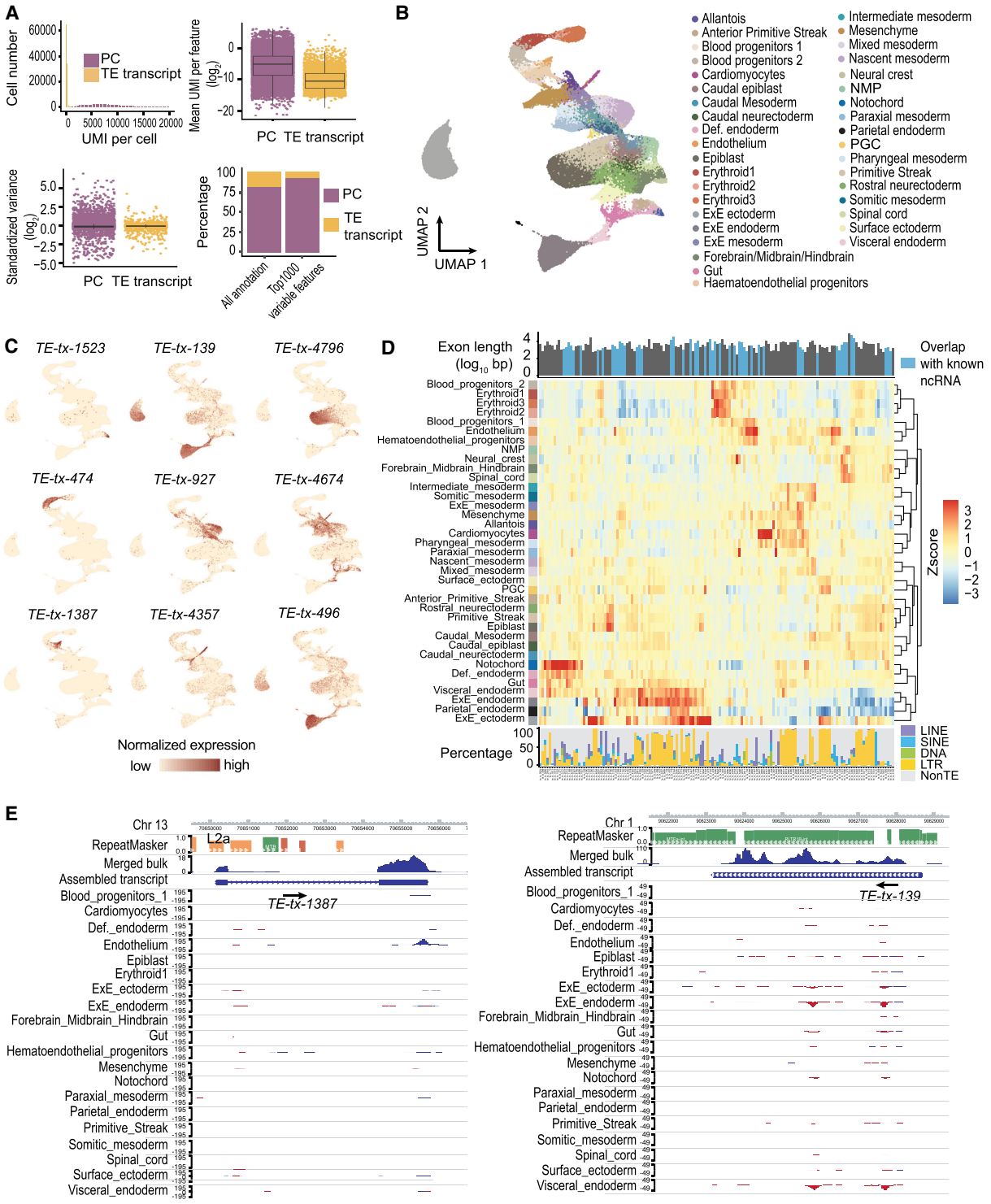

Figure 4. Tissue-specific TE expression during mouse gastrulation and early organogenesis. (*A*, *upper left*) Fewer unique molecular identifiers (UMIs) were mapped to TE transcripts than to protein-coding genes. (*Upper right*) The averaged expression level of TE transcripts across all the cells was lower compared to protein-coding genes. (*Lower left*) TE transcripts lack the extreme standardized variance observed at protein-coding genes. (*Lower right*) TE transcripts account for 73 of the top 1000 variable features. (*B*) UMAP of scRNA-seq data. Cells were colored based on tissue information provided by the original study. (*C*) Examples of tissue-specific TE transcripts. (*D*) Normalized expression pattern (*center*, heatmap) of 146 TE transcripts (columns) across 37 tissues (rows). Transcript length, annotation status (*top*, bar plot), and TE composition (*bottom*, bar plot) were shown for each TE transcript. (*E*) Genome browser view of two TE transcripts with strong tissue enrichment. Assembled TE transcripts, uniquely mapped reads of merged bulk RNA-seq (from 37 samples that were used for transcript assembly), and scRNA-seq signal for selected tissues were shown. (*Left*) A TE transcript that is initiated from an L2a element, the second exon of this transcript is composed of non-TE sequences. (*Right*) A TE transcript that is almost exclusively composed of ERV sequences.

strategy has been widely adopted in bulk RNA-seq and inspired similar analyses with scRNA-seq data. However, we caution that compared with bulk RNA-seq, a higher percentage of scRNA-seq reads are mapped to TEs, making it challenging to identify bona fide TE expression. Moreover, quantifying signal at individual TEs or TE subfamilies/families leads to the false impression that transcripts originated from repetitive regions are mostly composed of a single TE or TEs from the same subfamilies/families. Although this is true for some well-studied examples such as full-length ERVs, in most other cases, TEs only contribute to fragments of the full-length transcript, and TEs from different families can be incorporated into the same transcript.

A major difference between the expression quantification of protein-coding genes and TEs is that the transcript structures of protein-coding genes are usually well annotated and readily available. Gene annotation guides expression analysis toward genomic regions with true signal and facilitates accurate expression quantification with scRNA-seq data. In this study, we showed that transcripts constructed from bulk RNA-seq can serve as references for TE-containing ncRNAs and improve the accuracy of TE expression analysis in scRNA-seq data generated across multiple sequencing platforms. In comparison to individual TEs or TE subfamilies/families, TE transcripts more accurately reflect the natural transcription units. These transcripts contain not only previously annotated TE transcription units, but also novel ncRNAs that are partially composed of TEs. Of the 5299 TE transcripts that we assembled, 98 closely resemble the well-studied transcription units of ERVs. These transcripts have >80% of their exonic sequences contributed by TEs. They start from 5′ LTR, transcribe through internal sequences, and end at 3′ LTR. In addition, we also obtained 104 TE transcripts that initiate from TEs and are within 100 bp away from FANTOM5 CAGE peaks (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). Of these, 84 transcripts were not previously annotated, highlighting the value of our approach in identifying TEs that potentially function as promoters.

Using our analytical pipeline, we dissected the expression dynamics of TE transcripts during mouse early embryogenesis and identified 146 TE-containing ncRNAs with strong tissue specificity. A close examination of these candidates revealed intricate interaction between TE transcripts and protein-coding genes. For instance, we were able to identify a ChIP-seq peak of regulatory factor X, 3 (RFX3) at the promoter region of the TE transcript *TE-tx-3856* (Supplemental Fig. S13A). RFX3 is a transcription factor essential for brain development (Baas et al. 2006; Benadiba et al. 2012; Magnani et al. 2015). Our observation that *TE-tx-3856* is highly expressed in mouse neuronal tissues suggests that this TE-containing ncRNA is a potential downstream target of RFX3. In another example, the TE transcript *TE-tx-3715* overlaps with sonic hedgehog (*Shh*), a secreted signaling molecule produced by the notochord (Placzek 1995; McMahon et al. 2003). The expression pattern of *TE-tx-3715* strongly resembles that of *Shh*, indicating that they are under the control of a common regulatory circuit (Supplemental Fig. S13B). In addition, we also captured TE transcripts *TE-tx-3178* and *TE-tx-2841*, both of which are strongly expressed in the epiblast (Supplemental Fig. S13C). Both candidates initiate from TEs and were previously annotated as pluripotent associated transcripts (*Platr10* and *Platr14*) (Bergmann et al. 2015). Earlier reports suggested that *Platr10* transcript physically interacts with the promoter of pluripotent transcription factor *Sox2*, whereas the depletion of *Platr14* alters the expression of differentiation- and development-associated genes in stem cells (Bergmann et al. 2015; Zhang et al. 2019). Our observation that *Platr10* and *Platr14* are expressed in the epiblast suggests that they may play similar roles during mouse early embryogenesis. Taken together, we dissected the dynamic TE expression during mouse early development and provided a curated list of promising TE candidates for future functional studies.

In summary, we established an effective TE quantification pipeline for scRNA-seq data and illustrated the dynamic TE expression during mouse early embryogenesis. In contrast to commonly used bulk RNA-seq tools that evaluate reads at single TEs or TE subfamilies/families, our pipeline emphasizes the importance of full-length TE transcript structures in scRNA-seq TE quantification. Furthermore, our work provides an initial set of TE-containing long ncRNAs during mouse early development, laying the foundation for future work on constructing a more comprehensive TE transcript database across distinct tissue types and developmental stages and encompasses different types of TE transcripts, such as small RNAs and nonpolyadenylated long ncRNAs (McCue and Slotkin 2012; Dumesic and Madhani 2014; Lanciano and Cristofari 2020). Additionally, exploring effective techniques for quantifying TE-derived intronic reads (Chung et al. 2019; Kong et al. 2019; Navarro et al. 2019), as well as developing isoform-specific quantification tools for TE protein-coding gene chimeras (Wang et al. 2016; Pinson et al. 2018; Attig et al. 2019; Jang et al. 2019) will further expand the TE analysis toolkit for scRNA-seq and greatly advance our knowledge on the expression and the function of TE transcripts.

Methods

Data processing of bulk RNA-seq data sets

Raw sequencing files were downloaded from NCBI Sequence Read Archive and EMBL-EBI ArrayExpress (Supplemental Table S1) and aligned to the mouse (mm10) or human (hg38) genomes using STAR (Dobin et al. 2013). To retain reads derived from repetitive regions, "--outFilterMultimapNmax" was set to 500. To facilitate downstream transcript assembly "--outSAMattributes" was set to "NH HI NM MD XS AS." After alignment, signal quantification at regions of interests was performed using featureCounts. See "Read assignments" for details.

Data processing of scRNA-seq data sets

scRNA-seq data generated with Smart-seq-derived protocols were processed and quantified using the same procedures as bulk RNA-seq data. scRNA-seq data generated with the other protocols were processed using zUMIs (Parekh et al. 2018) with the following modifications: (1) To retain reads derived from repetitive regions, "--outFilterMultimapNmax" was set to 500 during STAR alignment. (2) To quantify reads that were mapped to multiple locations or features, "allowMultiOverlap" and "countMultiMappingReads" were set to TRUE for function ".runFeatureCount." BAM files with cell barcode, UMI, and the name of overlapping features were reported. (3) A customized R script (R Core Team 2017) was used to process the BAM file generated in step 2. See "Read assignments" for details.

Constructing TE transcripts

Transcript assembly of each RNA-seq sample was performed using StringTie2 (Kovaka et al. 2019). "-j 2 -s 5 -f 0.05 -c 2" was used to improve the specificity of the assembly results. To generate the master reference file, assembled transcripts from multiple RNA-seq samples were merged using TACO with default parameters

(Niknafs et al. 2017). Transcripts shorter than 200 nt were excluded. Transcripts whose exons that overlapped with TEs but not the exons of RefSeq protein-coding genes were named as TE transcripts and used for TE expression quantification.

Read assignments

FeatureCounts (Liao et al. 2014) was used to obtain the features to which each of the reads was mapped. BAM files with read information (read name, UMI, cell barcode) and overlapping features were reported. An EM algorithm was implemented to redistribute reads that mapped to multiple features. During the initial round of assignment, the n numbers of features without uniquely mapped reads first receive fractions of reads inversely proportional to the total number of features (N) to which each read is mapped. The remainder ($1-n/N$) read is then assigned to the remaining features proportionally to the amount of uniquely mapped reads after normalizing for feature length. During the subsequent rounds of assignment, reads mapped to multiple locations are reassigned to all the features proportionally to the amount of reads each feature received in the previous iteration after normalizing for feature length. The algorithm stops when the maximum read change per feature is smaller than 1 or the number of iterations reaches 50, whichever comes first.

RNA-seq simulation

To evaluate the performance of the EM algorithm, we simulated RNA-seq reads using Polyester (Frazee et al. 2015). One hundred base pair paired-end stranded RNA-seq reads were simulated using the 692 TE transcripts assembled from mESC data sets as templates. These simulated reads have a mean fragment length of 250 bp and sequencing error rate of 0.5%. TE transcripts were simulated to have a mean read coverage of 5× with expression deviation from the mean between twofold and 100-fold. We performed eight rounds of simulation with four independent expression designs (two replicates for each design). To evaluate the performance of the EM algorithm, we aligned the simulated reads to the mm10 genome and calculated the amount of observed reads at each TE transcript.

Mouse early embryogenesis scRNA-seq data set analysis

Reads of the scRNA-seq data sets from mouse zygote to gastrulation (Smart-seq-derived protocols) were quantified at protein-coding genes (RefSeq annotation, $n=20{,}779$) and TE transcripts (assembled from bulk RNA-seq, $n=5299$). Only cells that had 200–18,000 features and <10% mitochondria reads were kept. To remove batch effect and visualize all three data sets in the same UMAP, data integration was performed using Seurat with the top 4000 variable features (Stuart et al. 2019). Cell type was determined using the stage information provided by the original studies, the expression patterns of cell type–specific markers, and Seurat clustering results.

UMIs of the 10x scRNA-seq data set from mouse gastrulation to early organogenesis were quantified at protein-coding genes (RefSeq annotation, $n=20{,}779$) and TE transcripts (assembled from bulk RNA-seq, $n=5299$). Sample_25 was removed owing to higher batch effect. Only cells that have more than 200 features and were annotated by the original study were kept. Cell type information provided by the original study was used for identifying tissue-specific markers. The 146 TE transcripts with strong tissue enrichment were obtained by combining and filtering Seurat-defined markers and customized markers. Seurat-defined markers were obtained by running "FindAllMarkers" with "only.pos = T, min.pct = 0.10" and selecting for TE transcripts with adjusted P-value < 0.05. Customized markers were obtained by identifying TE transcripts with at least 1 UMI in at least 10% of the cells in any tissue and selecting candidates that were expressed in, at most, three tissues. After combining Seurat-defined markers and customized markers, manual curation was performed to remove candidates that were highly expressed in a large number of tissues or with suboptimal transcript structures.

TE transcript clustering in mouse preimplantation stages

Because of the limited number of cells, scRNA-seq signals of each TE transcript across all the cells with the same developmental stage were averaged to reduce noise. TE transcripts were then grouped into six clusters using soft clustering (R package TCseq) based on their expression patterns across preimplantation stages.

TE subfamily/family enrichment analysis

For each TE-family, its enrichment was calculated using the following equation: The observed frequency of TEs belonging to this family in all candidates divided by the expected frequency of TEs belonging to this family in genomic regions that do not overlap with protein-coding genes. The significance for the observed frequency was calculated with Fisher's exact test and corrected for multiple testing with the Benjamini–Hochberg method. Only TE families with more than 20 members in the candidates and more than 100 members in the background were included in the figures. TE subfamily enrichment analysis was performed similarly. Only TE subfamilies that were significantly enriched in the candidates had more than 10 members in the candidates and more than 100 members in the background and were plotted.

Other statistical analysis and figure generation

All the statistical analyses and associated figures were done using R (R Core Team 2017). Genome browser view was generated with WashU epigenome browser (https://epigenomegateway.wustl.edu/). A browser session showing the expression patterns of TE transcripts during mouse gastrulation and early organogenesis is available (https://epigenomegateway.wustl.edu/browser/?sessionFile=https://raw.githubusercontent.com/wanqingshao/TE_expression_in_scRNAseq/master/datahub/Gottgen_eg-session--da9eced0-e71d-11ea-be04-31bc80338b33.json).

Publicly available data sets used in this study

Descriptions and accession IDs of all the data sets used in this manuscript are provided in Supplemental Table S1.

Software availability

A detailed description of our analysis framework and customized scripts used for this work are publicly available at GitHub (https://github.com/wanqingshao/TE_expression_in_scRNAseq) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Kara Quaid and Shuonan He for comments on the manuscript; we thank all the Ting Wang laboratory members for discussions and critical input. This work is supported by National

Institutes of Health grants R01HG007175, U24ES026699, U01CA 200060, U01HG009391, and U41HG010972.

References

Anwar SL, Wulaningsih W, Lehmann U. 2017. Transposable elements in human cancer: causes and consequences of deregulation. *Int J Mol Sci* **18:** 974. doi:10.3390/ijms18050974

Attig J, Young GR, Hosie L, Perkins D, Encheva-Yokoya V, Stoye JP, Snijders AP, Ternette N, Kassiotis G. 2019. LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. *Genome Res* **29:** 1578–1590. doi:10.1101/gr.248922.119

Baas D, Meiniel A, Benadiba C, Bonnafe E, Meiniel O, Reith W, Durand B. 2006. A deficiency in RFX3 causes hydrocephalus associated with abnormal differentiation of ependymal cells. *Eur J Neurosci* **24:** 1020–1030. doi:10.1111/j.1460-9568.2006.05002.x

Benadiba C, Magnani D, Niquille M, Morlé L, Valloton D, Nawabi H, Ait-Lounis A, Otsmane B, Reith W, Theil T, et al. 2012. The ciliogenic transcription factor RFX3 regulates early midline distribution of guidepost neurons required for corpus callosum development. *PLoS Genet* **8:** e1002606. doi:10.1371/journal.pgen.1002606

Bendall ML, de Mulder M, Iñiguez LP, Lecanda-Sánchez A, Pérez-Losada M, Ostrowski MA, Jones RB, Mulder LCF, Reyes-Terán G, Crandall KA, et al. 2019. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol* **15:** e1006453. doi:10.1371/journal.pcbi.1006453

Bergmann JH, Li J, Eckersley-Maslin MA, Rigo F, Freier SM, Spector DL. 2015. Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Res* **25:** 1336–1346. doi:10.1101/gr.189027.114

Bialkowska AB, Yang VW, Mallipattu SK. 2017. Krüppel-like factors in mammalian stem cells and development. *Development* **144:** 737–754. doi:10.1242/dev.145441

Boroviak T, Stirparo GG, Dietmann S, Hernando-Herraez I, Mohammed H, Reik W, Smith A, Sasaki E, Nichols J, Bertone P. 2018. Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development* **145:** dev167833. doi:10.1242/dev.167833

Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. 2018. Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. bioRxiv doi:10.1101/462853

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33:** 155–160. doi:10.1038/nbt.3102

Cheng S, Pei Y, He L, Peng G, Reinius B, Tam PPL, Jing N, Deng Q. 2019. Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. *Cell Rep* **26:** 2593–2607.e3. doi:10.1016/j.celrep.2019.02.031

Chung N, Jonaid GM, Quinton S, Ross A, Sexton CE, Alberto A, Clymer C, Churchill D, Navarro Leija O, Han MV. 2019. Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. *Mob DNA* **10:** 39. doi:10.1186/s13100-019-0180-5

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18:** 71–86. doi:10.1038/nrg.2016.139

Cowley M, Oakey RJ. 2013. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet* **9:** e1003234. doi:10.1371/journal.pgen.1003234

Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15:** 583. doi:10.1186/1471-2164-15-583

De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D. 2017. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49:** 941–945. doi:10.1038/ng.3858

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

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. *Nat Rev Genet* **20:** 417–431. doi:10.1038/s41576-019-0117-3

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Donnison M, Beaton A, Davey HW, Broadhurst R, L'Huillier P, Pfeffer PL. 2005. Loss of the extraembryonic ectoderm in *Elf5* mutants leads to defects in embryonic patterning. *Development* **132:** 2299–2308. doi:10.1242/dev.01819

Dumesic PA, Madhani HD. 2014. Recognizing the enemy within: licensing RNA-guided genome defense. *Trends Biochem Sci* **39:** 25–34. doi:10.1016/j.tibs.2013.10.003

Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla M-E. 2013. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat Struct Mol Biol* **20:** 332–338. doi:10.1038/nsmb.2495

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507:** 462–470. doi:10.1038/nature13182

Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13:** 283–296. doi:10.1038/nrg3199

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* **46:** 558–566. doi:10.1038/ng.2965

Frazee AC, Jaffe AE, Langmead B, Leek JT. 2015. *Polyester*: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31:** 2778–2784. doi:10.1093/bioinformatics/btv272

Gaidatzis D, Burger L, Florescu M, Stadler MB. 2015. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol* **33:** 722–729. doi:10.1038/nbt.3269

Garcia-Perez JL, Widmann TJ, Adams IR. 2016. The impact of transposable elements on mammalian development. *Development* **143:** 4101–4114. doi:10.1242/dev.132639

Ge SX. 2017. Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development. *BMC Genomics* **18:** 200. doi:10.1186/s12864-017-3566-0

Gerdes P, Richardson SR, Mager DL, Faulkner GJ. 2016. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol* **17:** 100. doi:10.1186/s13059-016-0965-5

Gifford WD, Pfaff SL, Macfarlan TS. 2013. Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol* **23:** 218–226. doi:10.1016/j.tcb.2013.01.001

Göke J, Lu X, Chan YS, Ng HH, Ly LH, Sachs F, Szczerbinska I. 2015. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16:** 135–141. doi:10.1016/j.stem.2015.01.005

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522:** 221–225. doi:10.1038/nature14308

Hadjiargyrou M, Delihas N. 2013. The intertwining of transposable elements and non-coding RNAs. *Int J Mol Sci* **14:** 13307–13328. doi:10.3390/ijms140713307

He J, Babarinde IA, Sun L, Xu S, Chen R, Wei Y, Li Y, Ma G, Zhuang Q, Hutchins A, et al. 2020. Unveiling transposable element expression heterogeneity in cell fate regulation at the single-cell level. bioRxiv doi:10.1101/2020.07.23.218800

Hendrickson PG, Doráis JA, Grow EJ, Whiddon JL, Lim J-W, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, et al. 2017. 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

Huang Y, Kim JK, Do DV, Lee C, Penfold CA, Zylicz JJ, Marioni JC, Hackett JA, Surani MA. 2017. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *eLife* **6:** e22345. doi:10.7554/eLife.22345

Hutchins AP, Pei D. 2015. Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. *Sci Bull (Beijing)* **60:** 1722–1733. doi:10.1007/s11434-015-0905-x

Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17:** 29. doi:10.1186/s13059-016-0888-1

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. 2017. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49:** 1502–1510. doi:10.1038/ng.3945

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, et al. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet* **51:** 611–617. doi:10.1038/s41588-019-0373-3

Jeong HH, Yalamanchili HK, Guo C, Shulman JM, Liu Z. 2018. An ultra-fast and scalable quantification pipeline for transposable elements from

next generation sequencing data. *Pac Symp Biocomput* **23:** 168–179. doi:10.1142/9789813235533_0016
Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31:** 3593–3599. doi:10.1093/bioinformatics/btv422
Jonsson ME, Garza R, Sharma Y, Petri R, Sodersten E, Johansson JG, Johansson PA, Atacho DA, Pircs K, Madsen S, et al. 2020. Activation of endogenous retroviruses during brain development causes neuroinflammation. bioRxiv doi:10.1101/ 2020.07.07.191668
Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470. doi:10.1371/journal.pgen.1003470
Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13:** R107. doi:10.1186/gb-2012-13-11-r107
Kigami D, Minami N, Takayama H, Imai H. 2003. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol Reprod* **68:** 651–654. doi:10.1095/biolreprod.102.007906
Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Liu P, et al. 2015. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17:** 471–485. doi:10.1016/j.stem.2015.09.011
Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, Haverty PM, Tong AJ, Blanchette C, Albert ML, et al. 2019. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat Commun* **10:** 5228. doi:10.1038/s41467-019-13035-2
Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20:** 278. doi:10.1186/s13059-019-1910-1
La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560:** 494–498. doi:10.1038/s41586-018-0414-6
Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. *Nat Rev Genet* **21:** 721–736. doi:10.1038/s41576-020-0251-y
Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2017. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res* **45:** e17. doi:10.1093/nar/gkw953
Liao Y, Smyth GK, Shi W. 2014. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656
Lu X, Sachs F, Ramsay L, Jacques PÉ, Göke J, Bourque G, Ng HH. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21:** 423–425. doi:10.1038/nsmb.2799
Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487:** 57–63. doi:10.1038/nature11244
Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161:** 1202–1214. doi:10.1016/j.cell.2015.05.002
Magnani D, Morlé L, Hasenpusch-Theil K, Paschaki M, Jacoby M, Schurmans S, Durand B, Theil T. 2015. The ciliogenic transcription factor Rfx3 is required for the formation of the thalamocortical tract by regulating the patterning of prethalamus and ventral telencephalon. *Hum Mol Genet* **24:** 2578–2593. doi:10.1093/hmg/ddv021
Maksakova IA, Mager DL. 2005. Transcriptional regulation of early transposon elements, an active family of mouse long terminal repeat retrotransposons. *J Virol* **79:** 13865–13874. doi:10.1128/JVI.79.22.13865-13874.2005
McCue AD, Slotkin RK. 2012. Transposable element small RNAs as regulators of gene expression. *Trends Genet* **28:** 616–623. doi:10.1016/j.tig.2012.09.001
McKerrow W, Fenyö D. 2020. L1EM: a tool for accurate locus specific LINE-1 RNA quantification. *Bioinformatics* **36:** 1167–1173. doi:10.1093/bioinformatics/btz724
McMahon AP, Ingham PW, Tabin CJ. 2003. Developmental roles and clinical significance of hedgehog signaling. *Curr Top Dev Biol* **53:** 1–114. doi:10.1016/s0070-2153(03)53002-2
Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12:** 1483–1495. doi:10.1101/gr.388902
Mohammed H, Hernando-Herraez I, Savino A, Scialdone A, Macaulay I, Mulas C, Chandra T, Voet T, Dean W, Nichols J, et al. 2017. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep* **20:** 1215–1228. doi:10.1016/j.celrep.2017.07.009
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562. doi:10.1038/nature01262
Navarro FC, Hoops J, Bellfy L, Cerveira E, Zhu Q, Zhang C, Lee C, Gerstein MB. 2019. TeXP: deconvolving the effects of pervasive and autonomous transcription of transposable elements. *PLoS Comput Biol* **15:** e1007293. doi:10.1371/journal.pcbi.1007293
Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. 2017. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods* **14:** 68–70. doi:10.1038/nmeth.4078
Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura M, Tokunaga Y, Nakamura M, Watanabe A, et al. 2014. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci* **111:** 12426–12431. doi:10.1073/pnas.1413299111
O'Neill K, Brocks D, Hammell MG. 2020. Mobile genomics: tools and techniques for tackling transposons. *Philos Trans R Soc Lond B, Biol Sci* **375:** 20190345. doi:10.1098/rstb.2019.0345
Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. 2018. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7:** giy059. doi:10.1093/gigascience/giy059
Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7:** 597–606. doi:10.1016/j.devcel.2004.09.004
Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M. 2018. A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174:** 391–405.e19. doi:10.1016/j.cell.2018.05.043
Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, et al. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566:** 490–495. doi:10.1038/s41586-019-0933-9
Pikó L, Hammons MD, Taylor KD. 1984. Amounts, synthesis, and some properties of intracisternal A particle-related RNA in early mouse embryos. *Proc Natl Acad Sci* **81:** 488–492. doi:10.1073/pnas.81.2.488
Pinson ME, Pogorelcnik R, Court F, Arnaud P, Vaurs-Barrière C. 2018. CLIFinder: identification of LINE-1 chimeric transcripts in RNA-seq data. *Bioinformatics* **34:** 688–690. doi:10.1093/bioinformatics/btx671
Placzek M. 1995. The role of the notochord and floor plate in inductive interactions. *Curr Opin Genet Dev* **5:** 499–506. doi:10.1016/0959-437X(95)90055-L
Poznanski AA, Calarco PG. 1991. The expression of intracisternal A particle genes in the preimplantation mouse embryo. *Dev Biol* **143:** 271–281. doi:10.1016/0012-1606(91)90077-G
Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC, et al. 2012. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30:** 777–782. doi:10.1038/nbt.2282
R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46:** 21–42. doi:10.1146/annurev-genet-110711-155621
Ristevski S, O'Leary DA, Thornell AP, Owen MJ, Kola I, Hertzog PJ. 2004. The ETS transcription factor GABPα is essential for early embryogenesis. *Mol Cell Biol* **24:** 5844–5849. doi:10.1128/MCB.24.13.5844-5849.2004
Rodriguez-Terrones D, Torres-Padilla ME. 2018. Nimble and ready to mingle: transposon outbursts of early development. *Trends Genet* **34:** 806–820. doi:10.1016/j.tig.2018.06.006
Rowe HM, Trono D. 2011. Dynamic control of endogenous retroviruses during development. *Virology* **411:** 273–287. doi:10.1016/j.virol.2010.12.007
Santoni FA, Guerra J, Luban J. 2012. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9:** 111. doi:10.1186/1742-4690-9-111
Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, Elorbany R, Rhodes K, Burnett J, Gilad Y, et al. 2020. Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation. *Sci Rep* **10:** 1535. doi:10.1038/s41598-020-58327-6
Semrau S, Goldmann JE, Soumillon M, Mikkelsen TS, Jaenisch R, van Oudenaarden A. 2017. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun* **8:** 1096. doi:10.1038/s41467-017-01076-4

Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. 2014. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv doi:10.1101/003236

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177:** 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

Sundaram V, Wysocka J. 2020. Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Philos Trans R Soc Lond B, Biol Sci* **375:** 20190347. doi:10.1098/rstb.2019.0347

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol* **269:** 276–285. doi:10.1016/j.ydbio.2004.01.028

Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A. 2018. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci* **115:** 12565–12572. doi:10.1073/pnas.1814589115

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516:** 405–409. doi:10.1038/nature13804

Wang T, Santos JH, Feng J, Fargo DC, Shen L, Riadi G, Keeley E, Rosh ZS, Nestler EJ, Woychik RP. 2016. A novel analytical strategy to identify fusion transcripts between repetitive elements and protein coding-exons using RNA-Seq. *PLoS One* **11:** e0159028. doi:10.1371/journal.pone.0159028

Whiddon JL, Langford AT, Wong CJ, Zhong JW, Tapscott SJ. 2017. Conservation and innovation in the DUX4-family gene network. *Nat Genet* **49:** 935–940. doi:10.1038/ng.3846

Yandım C, Karakülah G. 2019. Expression dynamics of repetitive DNA in early human embryonic development. *BMC Genomics* **20:** 439. doi:10.1186/s12864-019-5803-1

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res* **47:** e27. doi:10.1093/nar/gky1301

Zhang S, Wang Y, Jia L, Wen X, Du Z, Wang C, Hao Y, Yu D, Zhou L, Chen N, et al. 2019. Profiling the long noncoding RNA interaction network in the regulatory elements of target genes by chromatin in situ reverse transcription sequencing. *Genome Res* **29:** 1521–1532. doi:10.1101/gr.244996.118

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8:** 14049. doi:10.1038/ncomms14049

Zhou J, Chehab R, Tkalcevic J, Naylor MJ, Harris J, Wilson TJ, Tsao S, Tellis I, Zavarsek S, Xu D, et al. 2005. *Elf5* is essential for early embryogenesis and mammary gland development during pregnancy and lactation. *EMBO J* **24:** 635–644. doi:10.1038/sj.emboj.7600538

Received April 27, 2020; accepted in revised form November 24, 2020.