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

Wanqing Shao
Ting Wang

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs
Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data

Wanqing Shao\textsuperscript{1,2} and Ting Wang\textsuperscript{1,2,3}

\textsuperscript{1}Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA; \textsuperscript{2}Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri 63110, USA; \textsuperscript{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 transposition, 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; Hadijargyrou and Delilas 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 MULERV-1, 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 (Criscone 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...
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 (Góke et al. 2015; Ge et al. 2017; Boroviak et al. 2018; Brocks et al. 2018; Yandam and Karakulah 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-uniform 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; Selewka 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 (Griscone 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; Kołodziejczyk 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.
TE expression quantification with scRNA-seq data

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 can 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 Fenyo 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 FANTOMS 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; Mäcsko 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

Downloaded from genome.cshlp.org on January 7, 2021 - Published by Cold Spring Harbor Laboratory Press
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) FANTOMS 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 RLR16B_Mm. This TE transcript overlaps Pla14, a long ncRNA known to impact the mESC differentiation-associated genes. (Right) A TE transcript that initiates from RLR16T_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; Gerdès et al. 2016; Rodriguez-Terzones 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 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 (Pikal et al. 1984; Pouznanski 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 ORR181-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 Kruppel-like factors, GABPA, and ELFS (Ristevski et al. 2004; Donnison et al. 2005; Zhou et al. 2005; Białkowska 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 hematopoietic lineage including hematopoietic progenitors, endothelium, blood progenitors, and erythrocytes 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...