Transposable Element Expression and Sub-cellular Dynamics During hPSC Differentiation to Endoderm, Mesoderm, and Ectoderm Lineages

Isaac A. Babarinde[1,2,*], Xiuling Fu[1,*], Gang Ma[1,*], Yuhao Li[3], Mobolaji T. Akinwole[4], Xuemeng Zhou[1], Andrew P. Hutchins[1#]

[1]Department of Systems Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China.

[2]Laboratory of Inflammation and Vaccines, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, 518055 Shenzhen, China.

[3]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia Qld 4072, Australia.

[4]Laboratories for Biomembrane Research and Biotechnology, Department of Biochemistry, University of Ibadan, Ibadan, Nigeria.

*Co-first authors.

#To whom correspondence should be addressed: andrewh@sustech.edu.cn

Abstract

Transposable elements (TEs) are genomic elements that are found in multiple copies in mammalian genomes. TEs were previously thought to have little functional relevance but recent studies have reported TE roles in multiple biological processes, particularly in embryonic development. To investigate the expression dynamics of TEs during human early development, we used long-read sequence data generated from *in vitro* differentiation of human pluripotent stem cells (hPSCs) to endoderm, mesoderm, and ectoderm lineages to construct lineage-specific transcriptome assemblies and accurately place TE sequences in their transcript context. Our analysis revealed that specific TE types, such as LINEs and LTRs, exhibit distinct expression patterns across different lineages. Notably, an expression outburst was observed in the ectoderm lineage, with multiple TE types showing dynamic expression trajectories. Additionally, certain LTRs, including HERVH and LTR7Y, were highly expressed in hPSCs and endodermal cells, but these HERVH and LTR7Y sequences originated from completely different transcripts. Interestingly, TE-containing transcripts exhibit distinct levels of transcript stability and subcellular localization across different lineages. Moreover, we showed a consistent trend of increased chromatin association of TE-containing transcripts in germ lineage cells compared to hPSCs. This study suggests that TEs contribute to human embryonic development through dynamic chromatin interaction.

Key findings:

- Different loci of the same TEs are independently regulated in different cell states
- Ectoderm has the highest frequency of TE-containing transcripts
- The presence of TEs dynamically drives transcripts to different sub-cellular compartments in different cell states
- hPSCs have the least stable TE transcripts with the weakest TE chromatin association, highlighting loose hPSC chromatin and potential roles in cell differentiation

Main

TEs are genomic elements with multiple copies resulting from autonomous and non-autonomous duplication in genomes [1,2]. About half of the human genome consists of TEs of different types and properties[1,3,4]. Because of their repetitive nature, TEs were previously considered to have little functional importance[5,6]. However, recent studies have shown that TEs are functionally important in diverse biological processes[2,4], including transcription[7–9], post-transcription expression regulation[10,11], transcript processing and stability[10,12,13], chromatin regulation[14,15], development[16,17] and disease progression[18,19]. While TEs have been co-opted in normal developmental processes[20,21], TEs still pose a risk to genomic integrity[4,22]. Therefore, TEs have both positive and negative aspects[6]. Although TEs have been studied in pluripotent stem cells[21,23] and some terminally differentiated somatic tissues[18,24], their expression patterns and their roles in post-implantation human development and gastrulation have not been well explored.

The presence of multiple copies of TEs in the genome makes the investigation of their functions difficult[25,26]. This is especially the case in TE expression quantification based on short-read RNA-seq in which a TE fragment often cannot be uniquely mapped to a genomic location[27]. Thus, assembling transcripts from only short-reads is fraught with difficulty even in well-annotated species such as human[28,29], and attempting to include assembly of TE-containing transcripts is doubly difficult[27]. Previous efforts have considered global TE expression without consideration for the specific TE loci[18,30,31], nor with their transcript context. Efforts are now being made to investigate TE expression at specific loci[23,25], with an emphasis on understanding how TE sequences are spliced into transcripts. Locus-level expression quantification has been reported to be improved by transcript assembly[32]. Although both short-read and long-read can be used for transcript assembly[23,26], the quality of the transcript assembly based on long-read data is superior[23,33]. Therefore, transcript assembly based on long-read data would substantially benefit TE expression studies.

We have previously shown that TE sequences are richly expressed in hPSCs, including inside coding and noncoding transcripts, and their presence is associated with changes in transcript biophysical processes[10]. It is generally assumed that TEs are most active in the pluripotent stage of development and decline in somatic tissues. However, this has not been explored, particularly in the context of TE sequences inside transcripts. In this study, we investigated human early

development transcripts (hEDTs) assembled exclusively from long-read data to investigate TE expression dynamics in human early development (hED). We firstly differentiated hPSCs *in vitro* to the three germ layers; endoderm, mesoderm, and ectoderm, to mimic human gastrulation and conducted long-read RNA sequencing for each cell type. By integrating deepCAGE and polyA data, we showed that our transcripts are virtually complete from the 5' to the 3' ends of the transcripts. Using the assemblies, we showed that TE-containing transcripts were dynamically regulated across different loci and cell states. We next investigated the expression dynamics, subcellular localization, and transcript stability of the assembled hEDTs with a focus on the TE-containing transcripts and found that the stability and chromatin interaction of TE-containing transcripts increases during the differentiation process. Overall, this study presents the expression dynamics of TEs in early development and implicates TEs in cell state changes.

hPSC transcriptome assembled from long-read RNA-Seq data

We have previously described an hPSC-specific transcriptome based on a combined long-read and short-read pipeline[23]. However, here we relied exclusively on deeply sequenced long-read data to assemble transcripts. This strategy has the advantage that TEs will be placed into their exact transcript context based only on long-reads, which will improve the accuracy of TE annotation in transcripts. To this end, we first generated PacBio long-read RNA-sequence data for hPSCs. Consensus reads were generated from the PacBio subreads, and then processed to produce noiseless reads using a standard PacBio read preparation pipeline (see Supplementary Methods) (https://github.com/nf-core/isoseq). The noiseless reads were then mapped to the human genome using Minimap2[34]. Next, StringTie2[35,36] was used to assemble transcripts. After quality assessment, 33,375 hPSC transcripts, corresponding to 16,814 genes, were assembled.

Reference-based transcript assembly with long-read data are not completely error-free[28,37,38]. Errors might arise because of the long-read region-biased error[39], reference-induced error[40], or random RNA shearing during extraction. We, therefore, checked if our assembled transcripts were full-length transcripts supported by long-read data, as opposed to annotation-aided assembly. Indeed, the assembled transcripts were covered entirely by multiple consensus reads (**Fig. 1a**), suggesting that the transcripts were expressed in hPSCs. One advantage of transcript assembly

based on long-read data only is the ability to sequence full-length mostly intact transcripts[26,27]. In support of this, most of the assembled transcripts were completely covered by at least one consensus read (**Fig. 1b**), suggesting that full-length transcripts were assembled. Interestingly, almost all the consensus reads used for the assembly were used without trimming (**Extended Data Fig. 1a**). As a negative control, we generated a 'random GTF set' with the same number and structure as the assembled transcripts (see Supplementary Methods). This random GTF was not supported by our long-read data (**Extended Data Fig. 1b**). The pileup of consensus reads shows that the reads map from the transcription start sites (TSSs) to the transcription end sites (TESs) of the hPSC assembly (**Fig. 1c**). The pileup of short-read data equally validates end-to-end transcript coverage for the hPSC assembly, but not for random GTF coordinates (**Extended Data Fig. 1c)**. These data highlight the power of long-reads in complete transcript assembly.

To further assess the completeness of the assembled transcripts, we utilized published hPSC deepCAGE[41] and polyadenylated (polyA) data that marks the TSS and TES, respectively. The long-read assembled TSSs were enriched for deepCAGE signal tags, while the TESs were enriched for the polyA signal (**Fig. 1d**). These enrichments were not found in the random GTF set (**Extended Data Fig. 1d)**. Similarly, ChIP-Seq data of hPSC POL2R, H3K27ac and H3K4me3 showed the expected promoter enrichment in the hPSC assembly, but not in the random GTF (**Fig. 1e, Extended Data Figs. 1e, f)**. Additionally, H3k36me3 signal is enriched on the hPSC transcript bodies (**Extended Data Fig. 1g**). The specific RNA-seq and histone modification data extensively validate the completeness of the hPSC assembly.

We next checked if the splice junctions of the assembled transcripts were supported. The evolutionary conservation scores across the junctions were higher relative to introns. The highest conservation was in the dinucleotide at the splice junctions for both the 5' and 3' ends (**Fig. 1f**). Indeed, the nucleotide frequencies showed the typical GT dinucleotide at the donor site, and AG at the acceptor site. Analysis of short-read RNA-seq data with an anchor length of 10 bp, showed that more than 99% of the splice junctions have at least one short-read junction support, and 85% have 10 or more reads (**Fig. 1g, Extended Data Fig. 1h)**. Taken together, these results, from multiple data, demonstrate the fidelity of the terminals and the splice junctions of the assembled transcripts.

In vitro differentiation drives consistent and specific transcriptome trajectory in germ layers

To investigate transcript expression dynamics during human embryonic development, we performed *in vitro* differentiation of the hPSC to three germ layers using previously published protocols[42] (**Fig. 2a**). Lineage-specific marker genes were specifically up-regulated in each lineage (**Extended Data Fig. 2)**. Long-read sequencing was done for two technical replicates for each cell state, and each sample had 20-45 million PacBio raw subreads (**Extended Data Fig. 3a)**. The consensus reads generated from each replicate ranged from 509 thousand to 1.2 million. Individual assembly produced 31,245 endoderm, 37,161 mesoderm, and 56,063 ectoderm transcripts (**Fig. 2b**). These corresponded to 14,989 endoderm, 17,520 mesoderm, and 24,766 ectoderm genes (**Extended Data Fig. 3b**). Since the number of assembled transcripts can potentially be affected by the sequencing depth[27], we asked if the higher transcript count in ectoderm cells is a genuine phenomenon or a consequence of greater sequencing depth in these samples. Interestingly, transcript assembly with individual ectoderm samples produced more transcripts even though the sequencing depths were lower than the combined sequencing depths in other cell states (**Extended Data Figs. 3a-b)**, suggesting that transcript complexity is higher in ectoderm cells. To further confirm this, we sampled 300,000 consensus reads from the replicates of each cell state. Transcript assembly from the uniform read counts revealed that more transcripts were assembled in ectoderm (**Extended Data Figs. 3c**). As the impact of sequencing depth tends to be more pronounced in lowly expressed transcripts[27], we checked if the higher transcript count in ectoderm would persist at a higher expression threshold. Indeed, that was the case up to 20 counts per million (CPM) expression threshold (**Extended Data Figs. 3d**), supporting the increased complexity of ectoderm transcripts.

Transcript quantification using the merged transcript reference showed that samples from the same cell state tended to cluster together and separated from other samples (**Fig. 2c, Extended Data Fig. 3e)**. Importantly, the differentiated samples were all separated from hPSC samples, and did not cluster with samples from other cell states, suggesting that different cell states had a specific transcriptomic landscape. Indeed, previously generated long-read data of H1 and c11 cell lines[23] consistently clustered with hPSC data, demonstrating the robustness of our pipelines to technical replicates and different hPSC cell lines. These results demonstrated the reliability of our *in vitro* differentiation experiments and transcript assembly and quantification procedures.

We then proceeded to investigate the features of hEDT across different cell types. Ectoderm transcripts tended to have longer spliced transcripts with fewer exons per transcript (Mann-Whitney U p value $< 1 \times 10^{-5}$) (**Figs. 2d, e**). However other parameters, including un-spliced transcript lengths, exon lengths and numbers of transcripts per gene were not substantially different across different cell states (**Extended Data Fig. 3f)**. We next predicted coding and noncoding potential using FEELnc[43], and transcript distributions based on protein coding potential showed that ectoderm tended to have more noncoding transcripts and expressed more noncoding RNA reads (**Fig. 2f**). The comparisons of the assemblies of each cell states to the GENCODE (v42) transcript set[44] revealed that ectoderm indeed had a lower proportion of GENCODE-known (matching) transcripts (**Extended Data Fig. 3g)**. Noncoding transcripts tended to be novel transcripts in all cell types, with a higher proportion of novel transcripts in ectoderm (**Fig. 2g**). Interestingly, base-level conservation levels, nucleotide frequencies of the splice junctions and the distributions of alternative splicing events were similar across each cell state (**Extended Data Figs. 3h-j)**.

To investigate the heterogeneity of transcript expression at the single cell level, previously published scRNA-seq data[45] containing hPSCs, and differentiated cell types was reanalyzed against our hEDT assembly. UMAP clustering revealed that cells of similar states tended to cluster together (**Extended Data Fig. 4a**). Meanwhile, hEDTs specifically expressed in each cell state and the top 20 uniformly expressed transcripts were extracted from the long-read data (**Fig. 2h**). We then checked the expression of the selected hEDTs and found that the expression patterns were consistently reflected in scRNA-seq cell states (**Fig. 2i, Extended Data Figs. 3b-f)**. Both long-read and scRNA-seq data largely consistently showed that the *in vitro* differentiation and bioinformatics procedures reliably captured the expressed hEDT set.

Different TE types are dynamically spliced into coding and noncoding transcripts during *in vitro* differentiation

Having established the reliability of the *in vitro* differentiation experiments and the assembled transcripts, we then investigated TE splicing patterns across the cell states. We used *nhmmer* to annotate TEs inside transcript sequences based on a previously described approach[2,46]. The

proportions of TE-containing transcripts and RNA read in the assembled transcripts revealed that ectoderm expressed more TE-containing transcripts (**Extended Data Fig. 5a**). As previously reported[23], most noncoding transcripts tended to contain more TEs than the coding transcripts (**Figs. 3a, b)**. This observation was true when we considered the number of TE-containing transcripts and total TE-containing RNA count, highlighting the coding potential disruption ability of TEs. Notably, in both coding and noncoding transcripts, ectoderm had more TEs. Intriguingly, most TE-containing transcripts contained less than half TE-derived sequence, and only a small proportion of the transcripts contained >50% TE-derived sequences (**Figs. 3c**). Again, there was a stark difference between coding and noncoding transcripts. TE-containing noncoding transcripts contained more TE-derived sequences, although most noncoding transcripts still contained less than 50% TE-derived sequences. The distributions of TE coverage were more similar in coding transcripts than in noncoding transcripts. Specifically, ectoderm TE-containing noncoding transcripts tended to contain more intermediate values. However, hPSC and endoderm contained the highest proportion of transcripts consisting entirely of TE-derived sequence (**Fig. 3c**) even though they had fewer overall TE-containing transcripts (**Fig. 3a**).

To investigate the specific TE splicing pattern further, we checked which TE types were contained in hEDTs. We found that the most frequently spliced TE types included SINE, retroposon, LINE, LTR, and DNA (**Fig. 3d**). It is important to note that these TE frequencies did not reflect the genomic TE distribution (**Fig. 3e**), suggesting the TEs are not just randomly spliced from genomic sequences. Across the major TE types, frequent TE splicing in ectoderm was observed with almost half of the transcripts containing at least one SINE, compared to about a third in hPSC transcripts (**Fig 3d**). Indeed, the higher frequency of SINE and retroposon splicing is observed across cell states in both coding and noncoding transcripts (**Extended Data Figs. 5b, c)**. On the contrary, less than 10% of the coding transcripts contained LTR-containing transcripts. The coding bias of the TEs was evident as the proportion of coding transcripts varied across cell states and TE types, ranging from 33% for LTR in ectoderm to 63% for SINE in endoderm (**Extended Data Fig. 5d**). While most of the TE-free transcripts were either known GENCODE transcripts (matching) or new isoforms of known GENCODE transcripts (variants), TE-containing transcripts, especially the noncoding transcripts, contained more novel transcripts (that do not overlap any known GENCODE transcript) with the highest novel proportion (~60%) for DNA TE-containing transcripts in ectoderm noncoding transcripts (**Extended Data Fig. 5e**). As TEs could

potentially create a minor transcript as an alternative isoform of a major functional transcript, we investigated if TE-containing transcripts tended to contain fewer major transcripts (transcripts with the highest expression among the isoforms of a single gene). For coding transcripts, the TE-free sequences tended to have a higher proportion of major transcripts, compared to the TE-containing transcripts (**Extended Data Fig. 5f**). Conversely, TE-containing noncoding transcripts tended to contain a higher proportion of major transcripts, suggesting that many TE-containing coding transcripts were the products of alternative splicing. The data demonstrated that our long-read assembly uncovered multiple previously unannotated transcripts and the bias of the insertion of various TE types into transcripts.

TEs might contribute to cell states or serve as functional modules for cell state transitions[47,48]. As previously reported[49], noncoding transcripts had higher expression variability, as quantified by the coefficient of variation, across cell types (**Fig. 3f**). Further, TE-containing coding and noncoding transcripts tended to have higher expression variability than TE-free transcripts. Interestingly, the Tau coefficient which measures expression specificity showed that both TE-containing coding and noncoding transcripts had higher cell-specific expression than TE-free transcripts, suggesting that TE expression is regulated in a cell type-specific manner. We next checked the expression variability within cell states using scRNA-seq data and found that hPSC transcripts had lower expression variability than differentiated cells (**Fig. 3g**). Except for mesoderm LTR-containing transcripts, noncoding transcripts tended to have higher expression variability than the coding transcripts. Notably, TE presence led to higher expression variability in coding transcripts but lower variability in noncoding transcripts. The differences in the expression variability might be influenced by the expression detectability, especially for lowly expressed transcripts. We checked the expression detectability and found that TE-containing transcripts tended to have higher detectability (**Fig. 3h**). Interestingly, while the transcript detectability in TE-free noncoding transcripts was lower than that of the TE-free coding transcripts, the detectability of the TE-containing coding and noncoding transcripts were more similar. These data showed that the low expression variability in TE-containing noncoding transcripts was not just because of the inability to detect due to the low transcript expression. Taken together, these results suggest that the trajectory of TE expression is coordinated during differentiation, but cell-level coding and noncoding transcript expression is dynamic.

Biased TE splicing patterns and frequencies of TE sequences in coding and noncoding transcripts

Having established the dynamic expression of TE-containing transcript in different lineages, we investigated the splicing patterns and insertion frequencies of different TEs in the four cell states. Because of the differences in coding and noncoding transcripts, we investigated the splicing patterns separately. Analysis of TE-splicing patterns in coding transcripts showed that TE sequences are rare in coding sequence (CDS) regions (**Fig. 4a, Extended Data Fig. 6a**), presumably reflecting the evolutionary cost of a TE inserting into and disrupting a coding sequence. Across TE types and cell states, the overrepresentation of TEs in the 5' and 3' untranslated regions (UTRs) was high, compared to the CDS, with higher frequencies in the 3' UTRs. However, investigation of specific TEs showed differences between different regions and cell states. For example, while HERVH and MSTB are both LTRs, their TE splicing patterns were different, with HERVH sequences being rare in the 3'UTR, whilst MSTBs were enriched (**Fig. 4a**). In addition, there was a higher frequency of HERVH TE-derived sequences in endoderm cells. Interestingly, the TE pattern of X24_LINE was very similar to that of MSTB, and also showed higher splicing in the UTRs of ectoderm, revealing the heterogeneity of the splicing patterns of specific TEs. The comparison of the MER3 and LTR6A frequencies further highlighted differences between these TEs and different cell states (**Extended Data Fig. 6a**).

We next looked at splicing at the TSS and TES boundaries of transcripts, and interestingly, while the TE splicing patterns across the TSS and TES were different, there were no substantial differences across cell states (**Fig. 4b**). Across the TE types, TE proportion was low at the TSS. In contrast, TEs were also reduced at the TES, although this reduction extended 5' into the transcript body (**Extended Data Fig. 6b**). Similarly, TE frequencies at both the donor and acceptor splice junctions were low (**Fig. 4c, Extended Data Fig. 6c**). Although there are subtle differences across cell lines and TE types, the general patterns were similar (**Extended Data Fig. 6c**). Surprisingly, TE frequencies of the exons at the splice sites suggest that TE insertion suppression at the exon starts is stronger than the suppression at the exon ends. This supports the idea that TEs are impaired from splicing into TSSs, TESs, or splice junctions directly to prevent the disruption of important motifs for transcription initiation and termination and transcript splicing.

We next studied the properties of TEs in noncoding transcripts and found that TE splicing patterns in noncoding transcripts were dynamic between TE types, cell states, and even positions within the transcript (**Fig. 4d, Extended Data Fig. 6d**). Specifically, while SINE and retroposon TEs are enriched towards the 5' and 3' ends of the transcripts, LINE, LTR, and DNA TEs are uniformly distributed. Additionally, the overall frequencies were different across cell states (**Fig. 4d, Extended Data Fig. 6d**). Ectoderm noncoding transcripts were enriched for LINE and DNA TEs, whilst LTRs were enriched in hPSCs, and were also high in ectoderm and endoderm (**Fig. 4d**). The dynamics of TE expression was more obvious for transcripts containing LTRs. For example, while HERVFH21 was concentrated in the center of noncoding transcripts, endoderm LTR6A had higher enrichment around 5' and 3' ends, suggesting these transcripts are noncoding remnants of intact ERVs. HERVH, conversely, was uniformly distributed across the noncoding transcripts in hPSC and endoderm cells (**Extended Data Fig. 6d**), highlighting the differences among TEs.

TE frequencies around noncoding transcript TSSs were also dynamic. SINE, LTRs, and retroposon were rare at the TSS (**Fig. 4e, Extended Data Fig. 6e)**. Conversely at the transcript ends, SINEs and LTRs were enriched up to the TES, and then their enrichment dropped substantially, suggesting that they marked transcript ends. TE sequences were absent at the TSSs of noncoding transcripts, suggesting TE sequences rarely act as core promoters of TSSs. TE frequencies at noncoding transcript splice junctions were also low (**Fig. 4e, Extended Data Fig. 6f)**, but not as low as in coding transcripts (**Extended Data Fig. 6f)**. Across the noncoding transcript bodies, TSS, TES, and splice junctions (**Figs. 4d-f, Extended Data Figs. 6d-f**), splice pattern differences based on the TE type and cell states were obvious.

To figure out the cell type-specific TE activities, we explored their frequencies in coding and noncoding transcripts. The comparisons of the frequencies of TE-containing coding transcripts identified 218 TEs that were significantly different in their frequencies upon differentiation into endoderm, mesoderm, or ectoderm (**Fig. 4g**). The majority of the TEs that were significantly different were enriched in ectoderm cells and were LINEs and LTRs and a few DNA TEs. In contrast, some LTRs like LTR7, HERVH, HERVH48, and MTSB2 were enriched in endoderm coding transcripts. On the other hand, the frequency comparison for noncoding transcripts identified 331 specific TEs that were differentially enriched in at least one of the four

cell states (**Fig. 4h**). Similar to the coding transcripts, the majority of the differentially spliced TEs in noncoding transcripts were LINEs and LTRs, and they were more frequent in the ectoderm. Some LTRs such as LTR7B and HERVH were more frequent in hPSCs as previously reported[50,51], and here we observed many of them in endoderm cells, suggesting these are not wholly specific features of hPSCs. Indeed, LTR7B/Y/H and HERVH were enriched in endoderm cells (**Fig. 4h**), supporting the idea that HERVH expression extends to cell types beyond hPSCs. HERVH modulates 3-dimensional genome structure in hPSCs[52] and may also be performing this function in somatic cells. Interestingly, there was a substantial overlap in the significantly enriched TE types in any cell type for both coding and noncoding transcripts (**Fig. 4i**). Breaking this down by TE type, the overlap is high for LINEs, with just 21-specific for noncoding transcripts and 134 LINE types were enriched in both coding and noncoding. For example, although the frequencies were different, the overall abundance patterns for LINE L1MD_orf2 were similar for coding and noncoding transcripts (**Fig. 4j**). In contrast to LINEs, LTRs were different between coding and noncoding transcripts, and only 49 LTR-types were common. This was exemplified by the LTRs MSTB2 and LTR7Y which were specific to cell types **(Fig. 4j)**. Overall, these data indicate that the TE content in different lineages is highly dynamic, whilst the TE sequences in coding and noncoding transcripts are more similar, except for LTRs, which are cell type and transcript-type specific.

TE-containing coding and noncoding transcripts expressed from different genomic loci are independently regulated

Above, we mainly consider TEs at the family or type level. However, a significant advantage of long-read sequence data is that we can place TEs into their specific transcript context. Although there is a substantial overlap in the types of TEs that are found in coding and noncoding transcripts, the exact genomic loci or the TE-containing transcripts are likely to be different. Potentially there are two scenarios, the first, is that the same TE type is contained in the same transcript, or that the same TE type is expressed, but from different transcripts. To explore this, we first needed to understand the relationship between TE type and expression level at the individual transcript level. We used bulk short-read RNA-seq data for each lineage with the long-read based hEDT reference, as the dynamic range of expression is higher for short-read data than long-read data[27]. In all four

cell states, the expression levels of TE-free coding transcripts were significantly higher (T-test p-value < 0.05) than those of TE-containing coding transcripts (**Extended Data Fig. 7a)**. TEs inserted into the 5' UTRs resulted in transcripts with the lowest expression levels while 3' UTR insertion led to comparatively higher expression levels, but the levels were not as high as those of the TE-free transcripts.

We next investigated the aggregate expression patterns and found that the overall expression patterns for SINE, LTR, LINE, retroposon, and DNA coding transcripts were not substantially different across the four states (**Fig. 5a**). However, differential transcript expression between hPSC and the three differentiated cell states revealed that hPSC-to-endoderm differentiation induced the least number of transcript changes while mesoderm differentiation drove the largest coding transcript expression changes (**Extended Data Fig. 7b)**. Differentiation in all three lineages led to both upregulation and downregulation of TE-containing coding transcripts, suggesting that transcript regulation is locus-specific and that TE expression is not restricted to hPSCs. We checked the overlaps of the differentiation-induced differential transcripts and found that many transcripts were regulated in a lineage-specific manner (**Extended Data Fig. 7c)**. The aggregated expression of specific TE coding transcripts identified seven LTR TEs (LTR7Y, LTR7C, LTR7B, LTR7, HERVH, HERVH48 and HERVFH21) that were significantly downregulated in both ectoderm and mesoderm cells, and two LINE TEs (L1P4b_5end and L1HS_5end) that were substantially upregulated in ectoderm cells (**Fig. 5a**, **Extended Data Figs. 7d, e**). This indicates that although the TE types are similar in all three cell types, they originate from different transcripts.

We then examined the expression levels of the individual TE-containing coding transcripts. Surprisingly, the majority of the HERVH-containing differentially expressed coding transcripts were more highly expressed in endoderm cells but not in hPSCs (**Fig. 5b**). Different coding transcripts containing L2 and L1HS_5end were activated in different cell states, with activation of many L1HS_5end-containing coding transcripts in ectoderm cells. These data support the second of our scenarios, at least for coding transcripts, that although similar TE types are expressed, they are derived from different loci and are expressed in cell type-specific transcripts.

As noncoding transcripts are more likely to be expressed in a cell type-specific manner[27,53,54], we expected that noncoding TE-containing transcripts would also be expressed from

different transcript loci. However, in comparison to the coding transcripts (**Extended Data Fig. 7b**), fewer differentially expressed noncoding transcripts were induced by the differentiation process (**Extended Data Fig. 7g**), reflecting the overall reduced number of noncoding transcripts in each cell type (**Fig. 2f**). Among the TE types, LTRs changed the most, with more LTR-containing noncoding transcripts downregulated in mesoderm cells. Also, many of the differentially expressed transcripts were regulated in a cell type-specific manner, and only a few differentially expressed transcripts were shared among cell states (**Extended Data Fig. 7h**). As for coding transcripts, in all cell states, TE-containing transcripts had lower overall expressions levels than TE-free transcripts (T-test p-value < 0.05) (**Extended Data Fig. 7i**). The comparison of the aggregate expression for noncoding transcripts containing specific TEs showed that 126 TEs were differentially expressed between the hPSC and at least one of the differentiated cell states (**Fig. 5c**). These TEs included multiple LTRs that are activated in hPSCs and endoderm cells. Interestingly, several SINE TEs were also differentially expressed. LTR6A and a few other LTR TEs were also found to be enriched for endoderm cells. The analyses of the differentially expressed LTR7-containing transcripts showed that many of the transcripts are downregulated in ectoderm and mesoderm cells, whilst LTR6A-containing noncoding transcripts were up in endoderm (**Fig. 5d**). Also, specific LINE-containing transcripts were specifically expressed in different lineages. Interestingly, half of the TEs with significant expression differences also had significantly different splicing frequencies (**Extended Data Fig. 7j**). These data support the model that similar TE types are expressed from different transcripts in different cell types.

We next investigated the expression dynamics of TE-containing coding transcripts at the single-cell level using scRNA-seq data from *in vitro* differentiated germ lineage cells[45]. The scRNA-seq data recaptured the higher expression of multiple TEs in ectoderm cells and downregulation of some LTRs such as LTR6A, LTR7, LTR7C, and HERVH in ectoderm and mesoderm cells (**Extended Data Fig. 8a**). The scRNA-seq data also revealed that the expression differences were not obvious at the level of TE types but became distinct at individual transcripts (**Fig. 5e, Extended Data Fig. 8b)**. Also, single-cell analyses of the noncoding transcripts were consistent with the observations from bulk data, and revealed upregulation of multiple TEs in ectoderm cells (**Extended Data Figs. 8a, b)**. Several LTR TEs were downregulated in ectoderm and mesoderm, while LTR6A was enriched in endoderm cells (**Fig. 5e, Extended Data Figs. 8c-d**). As with bulk data, lineage-specific expression became more obvious when TEs were

402 considered at the transcript level (**Fig. 5e, Extended Data Fig. 8d**). These results demonstrated
403 that the observations in bulk RNA-seq were recaptured in single cells.

404 To further elaborate on the expression dynamics of TE-containing transcripts we extended
405 our analysis to scRNA-seq from *in vivo* gastrulation data[55] (**Fig. 5f, Extended Data Fig. 9a**). As
406 expected, ectoderm cells were marked by more TE-containing coding and noncoding transcripts
407 (**Extended Data Figs. 9a, b)**. Indeed, compared to other cell types, more ectoderm-expressed
408 transcripts were found in almost all of the TE types in both coding and noncoding transcripts
409 (**Extended Data Figs. 9c)**. Also, many transcripts that were identified in the *in vitro* data were
410 recaptured in the *in vivo* data (**Fig. 5d, Extended Data Fig. 9d**), indicating that *in vitro*
411 observations reflected *in vivo* phenomena. Taken together, these data revealed that whilst similar
412 TE types are expressed in all cell types, they originate from multiple lineage-specific transcripts,
413 suggesting that TEs of the same type from multiple loci are independently regulated.

414

415 **TE-type switching of transcripts during differentiation into somatic cells**

416 As the same TE types are expressed from different transcripts, we next took advantage of
417 differentiation time course data to investigate how the expressions of TE-containing transcripts
418 vary during the endoderm differentiation process. As many LTR6A and HERVH-containing
419 transcripts are specifically expressed in the endoderm or hPSCs (**Figs. 5b, d)**, we focused our
420 initial analyses on the transcripts containing these two TEs. The analyses of the time-course bulk
421 RNA-seq data showed that the activation of LTR6A noncoding transcripts happened at 72 hours
422 in the differentiation process (**Figs. 6a, b)**. HERVH-containing noncoding transcripts with higher
423 endoderm expression were upregulated as early as 24 hours but did not reach maximum expression
424 until 72 hours. Conversely, HERVH-containing noncoding transcripts with high expression in
425 hPSCs were downregulated in endoderm at as early as 12 hours and were essentially undetectable
426 at 72 hours of differentiation to endoderm. The situation was similar for HERVH-containing
427 coding transcripts (**Figs. 6a, Extended Data Fig. 10a**). These results demonstrate how HERVH
428 and LTR6A can both be expressed in hPSCs and endoderm, but these TE sequences are contained
429 in lineage-specific transcripts.

Next, we analyzed scRNA-seq data of time-course hPSC-to-endoderm differentiation. Cell clustering by identity showed that the cells formed three distinct clusters representing early (0h to 24h) middle (36h) and late (72h and 96h) stages of differentiation (**Fig. 6c**). In agreement with the bulk RNA-seq analysis, endoderm-upregulated LTR6A, and HERVH-containing noncoding transcripts had high expression only at 72- and 96-hour of differentiation. In contrast, endoderm-downregulated noncoding transcripts were shut down at 12 hours of differentiation (**Fig. 6d**). These data showed that the timing for the activation and downregulation of TE transcripts varied during differentiation.

To further investigate if the expression of the endoderm-induced upregulated transcripts were maintained in terminally differentiated somatic cells, we first checked their expression in an array of hepatocyte-related samples. Surprisingly, the expression of noncoding transcripts with endoderm-induced upregulation was not sustained in hepatocyte-related terminally differentiated somatic cells (**Fig. 6e, Extended Data Fig. 10b, c**). Similar observations were found for HERVH-containing endoderm-upregulated coding transcripts (**Extended Data Fig. 10d**). Importantly, the expression of endoderm-specific HERVH-containing noncoding transcripts was not maintained in endoderm-derived terminally differentiated somatic cells (**Extended Data Fig. 11**). Conversely to HERVH, the expression of noncoding transcripts containing L1P1_orf2 and LTR that were specifically expressed in the ectoderm were maintained in multiple ectoderm-derived somatic cells, such as neurons (**Extended Data Fig. 11**). Similarly, the expression of many ectoderm-activated coding transcripts containing L2, L1P1_orf2 and LTR remained active in ectoderm-derived terminally differentiated somatic tissues. In summary, these results demonstrate how TE types can remain active across cell types, but they are often expressed from different transcripts in distinct tissues.

TE presence influences transcript subcellular localization

Previous studies have shown that TE-containing transcripts tend to localize in the nucleus[23,56]. To this end, we generated RNA-seq data for subcellular fractions of hPSCs, mesoderm, endoderm, and ectoderm differentiated cells for the nucleus and cytoplasm. Western blots for proteins known to localize to the nucleus and cytoplasm confirmed that the intended cellular fractions were purified

459 (**Extended Data Fig. 12a**). We confirmed our previous observation that both coding and
460 noncoding TE-containing transcripts preferentially localize to the nucleus in hPSCs[23] (**Fig. 7a**).
461 This pattern was the same for mesoderm and endoderm. Surprisingly, this was not the case for
462 ectoderm, where the opposite pattern was seen: TE-containing transcripts were instead more likely
463 to localize in the cytoplasm. Indeed, analyses of TE types revealed that they all tended to
464 preferentially localize to the cytoplasm in ectoderm cells, and to the nucleus in all other cell states
465 (**Extended Data Fig. 12b**). Greater details were revealed at the transcript level. For example, LTR-
466 containing noncoding transcripts tended to have different cytoplasm/nucleus enrichment when
467 compared to other TE types in PSC and endoderm (**Extended Data Fig. 12c**).

468 Noncoding RNAs that localize to the nucleus are often recruited to chromatin[57,58]. Hence,
469 we performed RNA-seq on the nucleoplasm and chromatin fractions to understand the sub-nuclear
470 localization of TE-containing transcripts (**Extended Data Fig. 12a**). TE-containing hPSC
471 transcripts were reduced in the chromatin compartment, versus the nucleoplasm (**Fig. 7b)**. On the
472 contrary, the TE containing transcripts of endoderm, mesoderm and ectoderm tended to be more
473 enriched in the chromatin fraction (**Fig. 7b, Extended Data Fig. 12d)**. Transcript-level analyses
474 of nucleoplasm/chromatin enrichment revealed heterogeneity across coding potentials, cell states
475 and TE types (**Extended Data Fig. 12e)**. These data reveal that the sub-cellular and sub-nuclear
476 localization of TE-containing transcripts is cell type-specific, and retention of TE-containing
477 transcripts in the nucleus is not a general feature.

478 To further investigate the subcellular transcript localization, we compared the expression
479 levels of different transcripts to obtain transcripts enriched in specific sub-cellular fractions (see
480 Supplementary methods). Pairwise comparisons of nucleus-cytoplasm and chromatin-
481 nucleoplasm expression patterns identified transcripts that were enriched in specific sub-cellular
482 fractions (**Extended Data Fig. 13)**. The number of transcripts localized to the sub-cellular
483 compartments varied across cell types (**Fig. 7c**). Importantly, compared to the differentiated cells,
484 hPSCs had fewer transcripts differentially localized to the nucleoplasm and chromatin. The
485 distribution of coding and noncoding transcripts revealed both cell state and location-specific
486 differences. Specifically, while nucleoplasm-localized transcripts had the highest noncoding
487 proportion in hPSC, chromatin-localized transcripts had the highest noncoding proportion in
488 differentiated cells (**Extended Data Fig. 14a**). Similarly, the distribution of TE-containing

transcripts varied across cell types, subcellular location, and protein-coding ability (**Extended Data Fig. 14b**). This effect was also transcript-specific, as few transcripts overlapped in the distinct sub-cellular compartments between the different cell types (**Extended Data Fig. 15**). Indeed, using a Jaccard similarity measure, transcripts that localized to the nucleus or cytoplasm were more consistent in hPSC, endoderm and mesoderm but more divergent in ectoderm (**Figs. 7d**). Interestingly, we found that the distribution of transcripts based on the number of cells in which they were localized varied across subcellular structures (**Fig. 7e**). The most unique chromatin localization was found in the mesoderm (**Figs. 7d**). Overall, nucleoplasm localization was the least consistent across the cell states. To further explore the relationship between transcript subcellular localization and cell state conversion, we checked the proportion of differentiation-induced DETs among the transcripts localized to various subcellular structures across cell states and found that subcellular localization significantly influenced the proportion of DETs (**Fig. 7f**), as the distribution of DETs were significantly different in localized transcripts compared to overall transcripts (Chi-square p-value < 0.05). For hPSCs, nucleoplasm-enriched transcripts had the lowest proportion of DETs. For differentiated cells, however, the lowest proportion of DETs was found in chromatin-localized transcripts. These data indicate that the localization of TE-containing transcripts is cell type-specific. Overall, compared to the TE-free transcripts, we found that TE-containing transcripts were more cell type-specific (**Fig. 7e**), with hPSCs having fewer chromatin-enriched TE transcripts.

TE presence influences transcript stability

As subcellular localization can influence degradation, we wondered if this would impact transcript half-life in a TE, sub-cellular localization, and cell state-dependent manner. We performed RNA-seq in differentiated cells treated with Actinomycin D to block transcription for 1 and 8 hours and measured the transcript levels relative to the 0-hour time-point. Surprisingly, there were divergent patterns of transcript half-life across TE types, coding potentials, and cell states (**Fig. 7g**). The lowest stability was found in hPSCs and endoderm. Interestingly, the impact of TE presence was also dynamic. While the TE-free transcripts were more stable in hPSC, ectoderm, and endoderm, this was not the case in mesoderm (**Fig. 7g, Extended Data Fig. 16a, b**). TE-containing ectoderm transcripts also contrasted with hPSCs and endoderm, as whilst TE-containing coding transcripts

were less stable, TE-containing noncoding transcripts were more stable than TE-free transcripts (**Fig. 7g**, **Extended Data Fig. 16a, b**). These revealed that transcript stability varied across cell states and transcript coding ability.

Finally, we checked if TE presence had any substantial influence on the stability of transcripts with differential subcellular localization. The stability at 8 hours after transcription stop showed differences across transcripts localized to different subcellular structures (**Fig. 7h**). While nuclear-localized transcripts were more stable than cytoplasm-localized transcripts in the ectoderm, the opposite was true in the other cell states. Conversely, chromatin-localized transcripts were more stable than nucleoplasm-localized transcripts in hPSCs, but the nucleoplasm-localized transcripts were more stable than chromatin-localized transcripts in differentiated cells. Interestingly, some differences in stability can be observed as early as 1 hour after transcription stop (**Extended Data Fig. 16c**). Moreover, the comparison of stabilities of TE-free transcripts to those containing TEs identified multiple TEs associated with significant differences in stability after 8 hours (**Extended Data Fig. 16d**). In hPSCs, TE presence mostly led to lower stability. In differentiated samples, mosaic patterns were found. For example, TE-containing transcripts localized to cytoplasm and nucleoplasm in mesoderm were more stable, along with cytoplasm-localized transcripts in ectoderm. These data suggest that TE presence might have different effects on transcript localization and stability across different cell states. Overall, the data suggest that stability and subcellular localization of TE-containing transcripts are dynamic across different cell state conversions

Discussion

TEs are difficult to place in their genomic context due to their repetitive nature. However, understanding expression dynamics would substantially benefit from robust TE-containing transcript assemblies[23,32]. Here, we used long-read RNA-seq technology to generate an accurate TE-containing transcript assembly[26,27]. These assembled transcripts enabled the accurate placement of TE sequences into transcripts in hPSCs and *in vitro*-derived germ layers. Our results showed that TEs are often expressed alongside unique genomic sequences. This partial overlap suggests that the TEs might have been inserted into a transcribed unit, relying on other

transcriptional machinery[1]. This possibility is supported by the evolutionary suppression of TE-sequences at the TSS and the relative overrepresentation inside the 3'UTRs of coding transcripts.

This study showed widespread and dynamic TE expression in human early development, although the frequencies and expression levels of different TE types vary between cell types. Multiple studies have shown the overexpression of TEs in hPSCs[7,8,59,60]. However, here we find that the expression of TE sequences is not restricted to hPSCs, and multiple TEs are expressed in both hPSCs and differentiated germ layers, Interestingly, although the same TE types are expressed, they are found in different transcripts. The ectoderm was particularly rich for TE-containing transcripts, and the expression of many TEs in ectoderm is sustained in the later stage of development, consistent with the reports of TE activity in terminally differentiated ectoderm-derived cells and tissues[30,61–63]. The exemptions to higher TE activities in ectoderm involved HERVH, LTR7, LTR6A, and several other LTRs that have previously been reported to be enriched in hPSCs[8,51]. In addition to expression in hPSCs, HERVH is also expressed in endoderm, and some HERVH-containing transcripts were higher in endoderm while others were higher in hPSCs. Curiously, many of the endoderm-expressed HERVH transcripts were not expressed in endoderm-derived somatic cells, again revealing transcript switching.

Consistent with previous studies [23,64,65], we found that TE presence leads to lower expression levels, probably due to the activities of RNA-binding proteins[13,65]. Indeed, SINE elements are overrepresented in 3' UTRs, and STAU1 promotes mRNA decay by targeting those Alu elements[13]. This indicates that the meta-analysis of TEs is limited, and TE sequences should be considered in their transcript context. From a biological perspective, it also suggests that cell-state specific regulation of TE expression occurs at the transcriptional stage or post-transcriptional stage. Interestingly, we found that TEs tend to be enriched in the chromatin fraction of the differentiated cells, suggesting that TE enrichment on chromatin might be associated with chromatin changes which might induce or regulate cell fate transitions. Taken together, this study reveals that TE-containing transcripts are highly dynamic in human early development, and the dynamics contribute to the cell state transitions by regulating chromatin structure in differentiating cells.

Acknowledgments

We acknowledge the assistance of SUSTech Core Research Facilities. Funding was from the National Natural Science Foundation of China (32270597, 31850410486) the Science Technology and Innovation Commission of Shenzhen (RCBS20221008093109033), and the Guangdong Basic and Applied Basic Research Foundation (2023A1515111170).

Accession data

Sequencing data from this study was deposited in the Gene Expression Omnibus with accession numbers GSE269270, GSE269272, GSE269273, and GSE269274.

Author contributions

I.A.B. planned and designed the study, performed most of the bioinformatic analyses, drafted the paper, and coordinated the experimental work. X.L. and G.M. performed the majority of the experiments. Y.L., M.T.A., and X.M. assisted with analysis. All authors assisted in writing the manuscript. A.P.H. designed the study, revised the manuscript, and supervised and funded the study.

Conflict of Interest

The authors declare no conflict of interest.

Methods

The detailed experimental procedure and bioinformatics pipelines for this study are presented in the Supplementary Methods.

References

601 1. Hutchins, A. P. & Pei, D. Transposable elements at the center of the crossroads between
602 embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. *Sci Bull*
603 *(Beijing)* **60**, 1722–1733 (2015).

604 2. Ma, G., Babarinde, I. A., Zhou, X. & Hutchins, A. P. Transposable Elements in Pluripotent Stem
605 Cells and Human Disease. *Front Genet* (2022).

606 3. Biémont, C. A brief history of the status of transposable elements: from junk DNA to major
607 players in evolution. *Genetics* **186**, 1085–93 (2010).

608 4. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 1–
609 12 (2018).

610 5. Kim, Y.-J., Lee, J. & Han, K. Transposable Elements: No More 'Junk DNA'. *Genomics Inform* **10**, 226–
611 33 (2012).

612 6. Naville, M. *et al.* Not so bad after all: retroviruses and long terminal repeat retrotransposons as a
613 source of new genes in vertebrates. *Clinical Microbiology and Infection* **22**, 312–323 (2016).

614 7. Wang, J., Huang, J. & Shi, G. Retrotransposons in pluripotent stem cells. *Cell Regeneration* **9**,
615 (2020).

616 8. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem
617 cells. *Nature* **516**, (2014).

618 9. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression.
619 *Science (1979)* **351**, (2016).

620 10. Babarinde, I. A. *et al.* Transposable element sequence fragments incorporated into coding and
621 noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids*
622 *Res* **49**, 9132–9153 (2021).

623 11. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of
624 mammalian transcription. *Nat Rev Mol Cell Biol* **23**, 481–497 (2022).

625 12. Farré, D., Engel, P. & Angulo, A. Novel Role of 3'UTR-Embedded Alu Elements as Facilitators of
626 Processed Pseudogene Genesis and Host Gene Capture by Viral Genomes. *PLoS One* **11**,
627 e0169196 (2016).

628 13. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3'
629 UTRs via Alu elements. *Nature* **470**, 284–290 (2011).

630 14. Raviram, R. *et al.* Analysis of 3D genomic interactions identifies candidate host genes that
631 transposable elements potentially regulate. *Genome Biol* **19**, (2018).

632 15. Sun, L., Fu, X., Ma, G. & Hutchins, A. P. Chromatin and Epigenetic Rearrangements in Embryonic
633 Stem Cell Fate Transitions. *Front Cell Dev Biol* **9**, 174 (2021).

634 16. Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on
635 mammalian development. *Development* **143**, 4101–4114 (2016).

636 17. Hackett, J. A., Kobayashi, T., Dietmann, S. & Surani, M. A. Activation of Lineage Regulators and
637 Transposable Elements across a Pluripotent Spectrum. *Stem Cell Reports* **8**, (2017).

638 18. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of
639 repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, (2014).

640 19. LaRocca, T. J., Cavalier, A. N. & Wahl, D. Repetitive elements as a transcriptomic marker of aging:
641 Evidence in multiple datasets and models. *Aging Cell* **19**, (2020).

642 20. Kruse, K. *et al.* Transposable elements drive reorganisation of 3D chromatin during early
643 embryogenesis. *bioRxiv* 523712 (2019) doi:10.1101/523712.

644 21. Nellåker, C. *et al.* The genomic landscape shaped by selection on transposable elements across 18
645 mouse strains. *Genome Biol* **13**, R45 (2012).

646 22. Gray, Y. H. M. It takes two transposons to tango:transposable-element-mediated chromosomal
647 rearrangements. *Trends in Genetics* **16**, 461–468 (2000).

648 23. Babarinde, I. A. *et al.* Transposable element sequence fragments incorporated into coding and
649 noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids*
650 *Res* **49**, 9132–9153 (2021).

651 24. Chishima, T. *et al.* Identification of Transposable Elements Contributing to Tissue-Specific
652 Expression of Long Non-Coding RNAs. *Genes (Basel)* **9**, 23 (2018).

653 25. Berrens, R. V. *et al.* Locus-specific expression of transposable elements in single cells with CELLO-
654 seq. *Nat Biotechnol* **40**, (2022).

655 26. Babarinde, I. A., Li, Y. & Hutchins, A. P. Computational Methods for Mapping, Assembly and
656 Quantification for Coding and Non-coding Transcripts. *Comput Struct Biotechnol J* **17**, (2019).

657 27. Babarinde, I. A. & Hutchins, A. P. The effects of sequencing depth on the assembly of coding and
658 noncoding transcripts in the human genome. *BMC Genomics* **23**, 487 (2022).

659 28. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid
660 of long and short reads with StringTie. *PLoS Comput Biol* **18**, (2022).

661 29. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with Capture
662 Long-Read Sequencing. *Nat Genet* **49**, 1731 (2017).

663 30. He, J. *et al.* Identifying transposable element expression dynamics and heterogeneity during
664 development at the single-cell level with a processing pipeline scTE. *Nat Commun* **12**, (2021).

665 31. Rodríguez-Quiroz, R. & Valdebenito-Maturana, B. SoloTE for improved analysis of transposable
666 elements in single-cell RNA-Seq data using locus-specific expression. *Commun Biol* **5**, (2022).

667 32. Shao, W. & Wang, T. Transcript assembly improves expression quantification of transposable
668 elements in single-cell RNA-seq data. *Genome Res* **31**, 88–100 (2021).

669 33. Bayega, A., Fahiminiya, S., Oikonomopoulos, S. & Ragoussis, J. Current and future methods for
670 mRNA analysis: A drive toward single molecule sequencing. in *Methods in Molecular Biology* vol.
671 1783 209–241 (Humana Press, New York, NY, 2018).

672 34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
673 (2018).

674 35. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq
675 reads. *Nat Biotechnol* **33**, 290–295 (2015).

676 36. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis
677 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650–67 (2016).

678 37. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly
679 evaluation with Inspector. *Genome Biol* **22**, (2021).

680 38. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction.
681 *Nature Biotechnology 2019 37:2* **37**, 124–126 (2019).

682 39. Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads.
683 *NAR Genom Bioinform* **4**, (2022).

684 40. Nip, K. M. *et al.* Reference-free assembly of long-read transcriptome sequencing data with RNA-
685 Bloom2. *Nat Commun* **14**, (2023).

686 41. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and
687 GRCm38 genome assemblies. *Sci Data* **4**, 170107 (2017).

688 42. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation.
689 *Nature* **518**, 344–9 (2015).

690 43. Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog
691 transcriptome. *Nucleic Acids Res* **45**, e57 (2017).

692 44. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023.
693 *Nucleic Acids Res* **51**, D942–D949 (2023).

694 45. Chu, L. F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell
695 differentiation to definitive endoderm. *Genome Biol* **17**, (2016).

696 46. Wheeler, T. J. & Eddy, S. R. Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**,
697 2487–2489 (2013).

698 47. Sakashita, A. *et al.* Transcription of MERVL retrotransposons is required for preimplantation
699 embryo development. *Nat Genet* **55**, (2023).

700 48. Modzelewski, A. J. *et al.* A mouse-specific retrotransposon drives a conserved Cdk2ap1 isoform
701 essential for development. *Cell* **184**, (2021).

702 49. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global
703 properties and specific subclasses. *Genes Dev* **25**, 1915–1927 (2011).

50. Santoni, F. A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**, (2012).

51. Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21**, (2014).

52. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, (2019).

53. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* **24**, (2023).

54. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res* **22**, 1760–1774 (2012).

55. Tyser, R. C. V. *et al.* Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature* **600**, (2021).

56. Mas-Ponte, D. *et al.* LncATLAS database for subcellular localization of long noncoding RNAs. *Rna* **23**, 1080–1087 (2017).

57. Yin, Y. *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, (2020).

58. Mishra, K. & Kanduri, C. Understanding Long Noncoding RNA and Chromatin Interactions: What We Know So Far. *Non-Coding RNA* **5**, 54 (2019).

59. Mustafin, R. N. The role of transposable elements in the differentiation of stem cells. *Molecular Genetics Microbiology and Virology (Russian version)* **37**, (2019).

60. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev* **28**, (2014).

61. Pereira Fernandes, D., Bitar, M., Jacobs, F. & Barry, G. Long Non-Coding RNAs in Neuronal Aging. *Noncoding RNA* **4**, 12 (2018).

62. Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. & Gage, F. H. The role of transposable elements in health and diseases of the central nervous system. *J Neurosci* **33**, 17577–86 (2013).

63. Lapp, H. E. & Hunter, R. G. Early life exposures, neurodevelopmental disorders, and transposable elements. *Neurobiol Stress* **11**, 100174 (2019).

64. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**, R107 (2012).

65. Kelley, D. R., Hendrickson, D. G., Tenen, D. & Rinn, J. L. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**, 537 (2014).

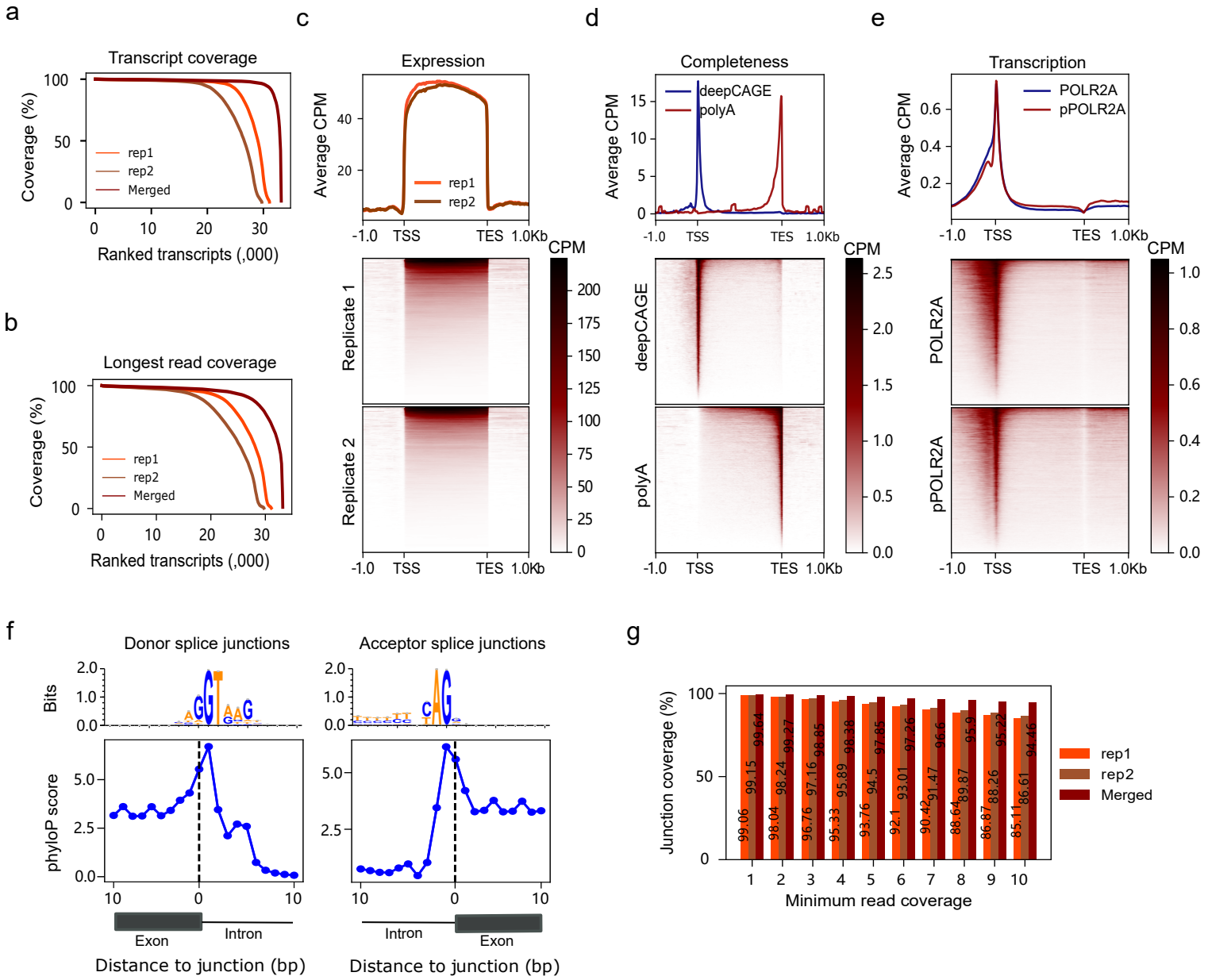
a
Transcript coverage
Coverage (%)
rep1
rep2
Merged
Ranked transcripts (,000)
b
Longest read coverage
Coverage (%)
rep1
rep2
Merged
Ranked transcripts (,000)
c
Expression
Average CPM
rep1
rep2
-1.0 TSS TES 1.0Kb
CPM
Replicate 1
Replicate 2
d
Completeness
Average CPM
deepCAGE
polyA
CPM
deepCAGE
polyA
e
Transcription
Average CPM
POLR2A
pPOLR2A
CPM
POLR2A
pPOLR2A
f
Donor splice junctions
Acceptor splice junctions
Bits
phyloP score
Exon
Intron
Distance to junction (bp)
g
Junction coverage (%)
rep1
rep2
Merged
Minimum read coverage

Fig. 1

Fig. 1: Quality assessment of the assembled hPSC long-read transcriptome. **a**, Consensus read coverage of the assembled PSC transcripts using all consensus reads. **b**, The longest consensus read coverage for the assembled PSC transcripts. **c**, Pileups for the consensus long reads for the assembled hPSC transcripts across the exons from the TSS to the TES. Each transcript is scaled to a uniform size. **d**, DeepCAGE (5' end) and polyA tail (3' end) pileups of the assembled transcripts, from the TSS to the TES scaled to a uniform size and not including intronic regions. PolyA data were from GSE138759 and GSE111134 while deepCAGE data were from GSE34448 and GSE61264. **e**, Density pileup of POLR2A and phosphorylated POLR2A (pPOLR2A) for the assembled hPSC transcripts. The data were from GSE242645. **f**, Position weight matrix nucleotide frequencies and phyloP conservation scores of the 10-bp positions around the donor and acceptor splice junctions of the assembled multi-exon hPSC transcripts **g**, Proportion of hPSC junctions covered by indicated minimum number of reads.

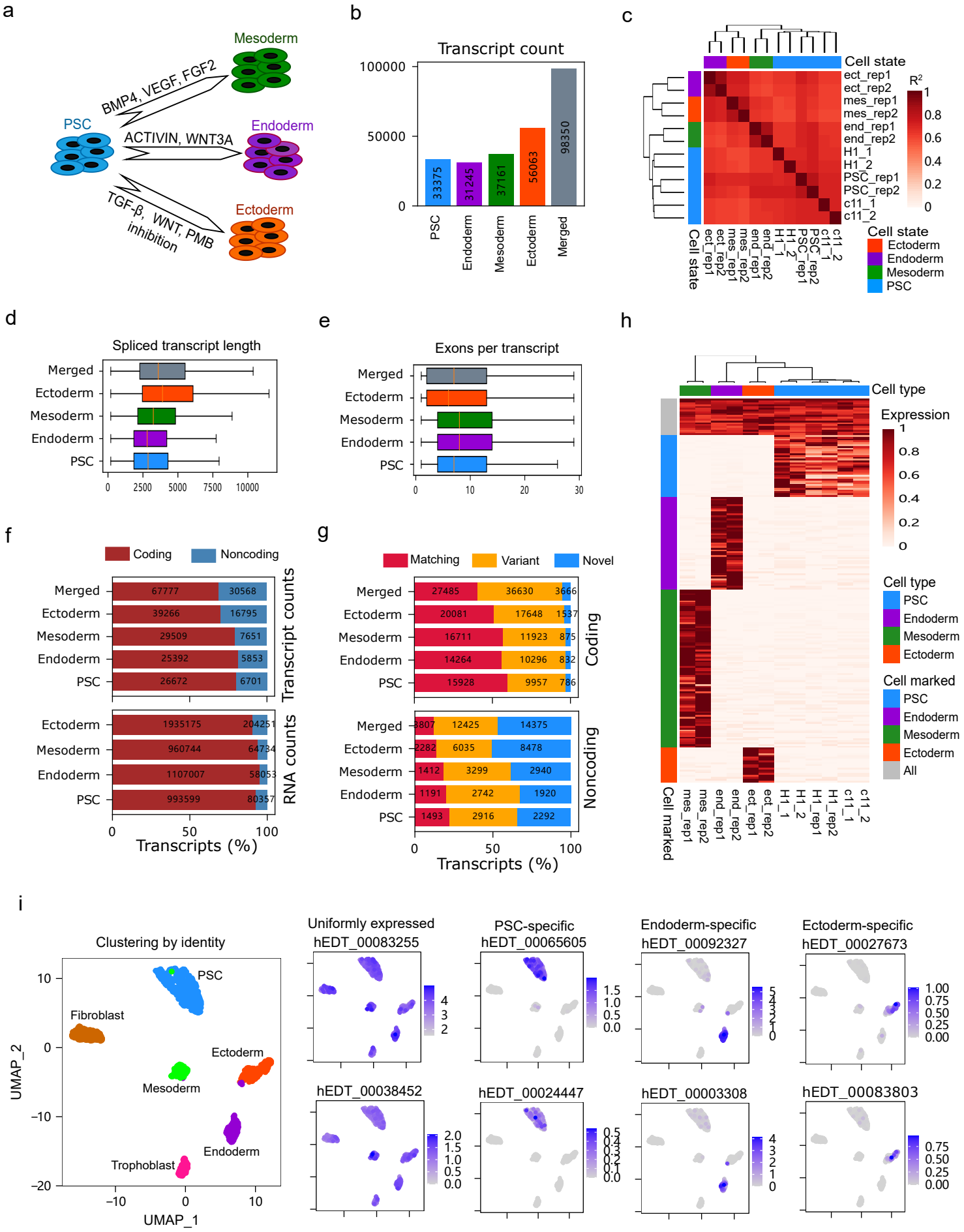
a
Mesoderm
BMP4, VEGF, FGF2
PSC
ACTIVIN, WNT3A
Endoderm
TGF-β, WNT, PMB inhibition
Ectoderm
b
Transcript count
100000
50000
0
33375
31245
37161
56063
98350
PSC
Endoderm
Mesoderm
Ectoderm
Merged
c
Cell state
ect_rep1
ect_rep2
mes_rep1
mes_rep2
end_rep1
end_rep2
H1_1
H1_2
PSC_rep1
PSC_rep2
c11_1
c11_2
R²
1
0.8
0.6
0.4
0.2
0
Cell state
Ectoderm
Endoderm
Mesoderm
PSC
d
Spliced transcript length
Merged
Ectoderm
Mesoderm
Endoderm
PSC
0
2500
5000
7500
10000
e
Exons per transcript
Merged
Ectoderm
Mesoderm
Endoderm
PSC
0
10
20
30
h
Cell type
Expression
1
0.8
0.6
0.4
0.2
0
Cell type
PSC
Endoderm
Mesoderm
Ectoderm
Cell marked
PSC
Endoderm
Mesoderm
Ectoderm
All
Cell marked
mes_rep1
mes_rep2
end_rep1
end_rep2
ect_rep1
ect_rep2
H1_1
H1_2
H1_rep1
H1_rep2
c11_1
c11_2
f
Coding
Noncoding
Merged 67777 30568
Ectoderm 39266 16795
Mesoderm 29509 7651
Endoderm 25392 5853
PSC 26672 6701
Transcript counts
Ectoderm 1935175 204251
Mesoderm 960744 64734
Endoderm 1107007 58053
PSC 993599 80357
RNA counts
0
50
100
Transcripts (%)
g
Matching
Variant
Novel
Merged 27485 36630 3666
Ectoderm 20081 17648 1537
Mesoderm 16711 11923 875
Endoderm 14264 10296 832
PSC 15928 9957 786
Coding
Merged 3807 12425 14375
Ectoderm 2282 6035 8478
Mesoderm 1412 3299 2940
Endoderm 1191 2742 1920
PSC 1493 2916 2292
Noncoding
0
50
100
Transcripts (%)
i
Clustering by identity
PSC
Fibroblast
Ectoderm
Mesoderm
Endoderm
Trophoblast
UMAP_2
UMAP_1
10
0
−10
−20
−10
0
10
Uniformly expressed
hEDT_00083255
PSC-specific
hEDT_00065605
Endoderm-specific
hEDT_00092327
Ectoderm-specific
hEDT_00027673
hEDT_00038452
hEDT_00024447
hEDT_00003308
hEDT_00083803

Fig. 2

751 **Fig. 2: Expression dynamics of the assembled human early development transcripts**. **a**,
752 Experimental design for the *in vitro* differentiation of human hPSC to each of the three embryonic
753 germ layers. **b**, The numbers of transcripts assembled from each cell state and the merged transcript
754 assembly. **c**, The expression correlation heatmap of the samples made from long-read
755 quantification. H1_1, H1_2, c11_1 and c11_2 samples were from previously reported data[23]. **d**,
756 Boxplots showing the spliced transcript lengths of the assembled transcripts. Boxplots show the
757 mean (red central bar), and second and third quartiles, and the whiskers show 1.5 times the
758 interquartile ranges, for this and all subsequent boxplots. Kruskal-Wallis p value = 2.37×10^{-28}.
759 **e**, Boxplots showing the number of exons per assembled transcript. Kruskal-Wallis p value = 1.86
760 $\times 10^{-173}$. **f**, Stacked bar chart showing the proportions and numbers of coding and noncoding
761 transcripts in the hEDT assemblies. The upper plot shows the distribution of the assembled
762 transcript counts while the lower plot shows the distribution of the total RNA counts from the long-
763 read quantifications. **g**, Stacked bar chart showing the proportions and numbers of assembled
764 transcripts based on the similarity to GENCODE assembly. Matching transcript completely
765 matches to a known transcript (including all exons and splice junctions), while a variant transcript
766 partially overlaps a known GENCODE transcript. Novel transcript does not overlap any known
767 GENCODE transcript exon. **h**, Heatmap showing the uniformly expressed transcripts and
768 transcripts specific to each cell state based on the long-read quantification. For each transcript, the
769 expression was normalized by the highest expression level. **i**, Single-cell RNA-seq expression
770 showing uniformly expressed transcripts and marker transcripts specific to the indicated cell states.

771

772

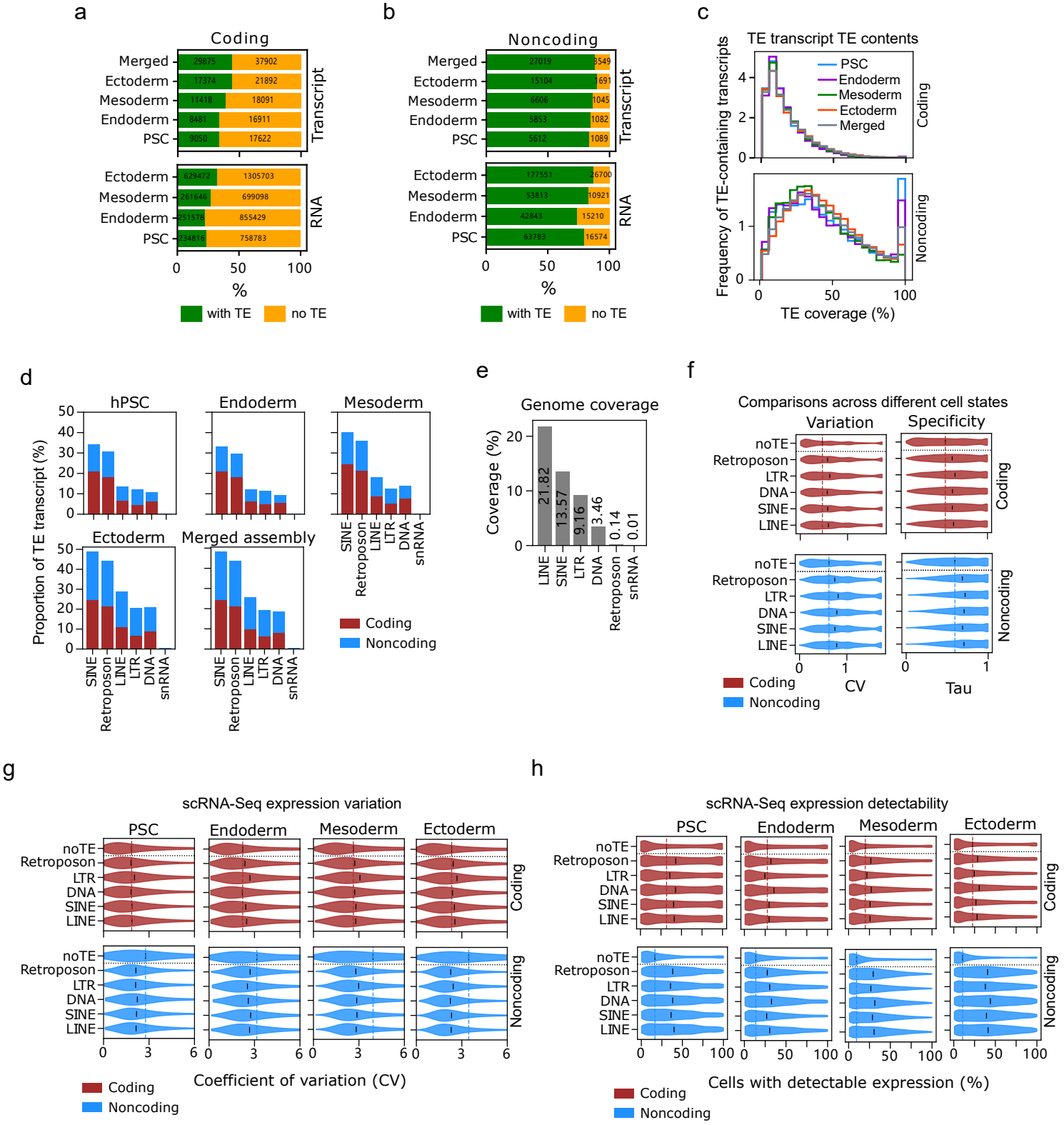
a
Coding
Merged
Ectoderm
Mesoderm
Endoderm
PSC
29875
37902
17374
21892
11418
18091
8481
16911
9050
17622
Transcript
629472
1305703
261646
699098
851578
855429
734816
758783
RNA
0
50
100
%
with TE
no TE
b
Noncoding
27019
3549
15104
1691
6606
1045
5853
1082
5612
1089
177551
26700
53813
10921
42843
15210
63783
16574
c
TE transcript TE contents
PSC
Endoderm
Mesoderm
Ectoderm
Merged
Coding
Noncoding
Frequency of TE-containing transcripts
TE coverage (%)
d
hPSC
Endoderm
Mesoderm
Ectoderm
Merged assembly
Proportion of TE transcript (%)
SINE
Retroposon
LINE
LTR
DNA
snRNA
Coding
Noncoding
e
Genome coverage
Coverage (%)
21.82
13.57
9.16
3.46
0.14
0.01
LINE
SINE
LTR
DNA
Retroposon
snRNA
f
Comparisons across different cell states
Variation
Specificity
noTE
Retroposon
LTR
DNA
SINE
LINE
Coding
Noncoding
CV
Tau
g
scRNA-Seq expression variation
PSC
Endoderm
Mesoderm
Ectoderm
Coefficient of variation (CV)
h
scRNA-Seq expression detectability
Cells with detectable expression (%)

Fig. 3

773 **Fig. 3: Divergent TE expression dynamics across coding and noncoding transcripts**. **a-b**,
774 Stacked bar charts showing the proportion and number of TE-containing transcripts in the coding
775 (**a**) and noncoding (**b**) transcripts in the hEDT assemblies. The upper plots show the transcript
776 counts while the lower plots show the RNA counts from long-read based transcript quantification.
777 **c**, Histograms showing the distributions of TE coverages (Percent of sequence that consists of TE)
778 for both coding and noncoding transcripts. **d**, Stacked bar charts showing the proportion of TE-
779 containing transcripts for coding and noncoding transcripts in the assembled hEDTs. **e**, Genomic
780 distributions of the top expressed TE types. **f**, Violin plots based on long-read quantification
781 showing the expression variability of TE-containing transcripts across different cell states. The left
782 charts show the expression variability measured by the coefficient of variation, while the right
783 charts show the expression specificity as measured by Tau. **g**, Single-cell RNA-seq expression
784 variability of different groups of transcripts within the cell states. **h**, Single-cell RNA-seq
785 expression detectability of different groups of transcripts within the cell states. The red and blue
786 dotted lines represent the median values for the coding and noncoding transcripts without TEs.

787

788

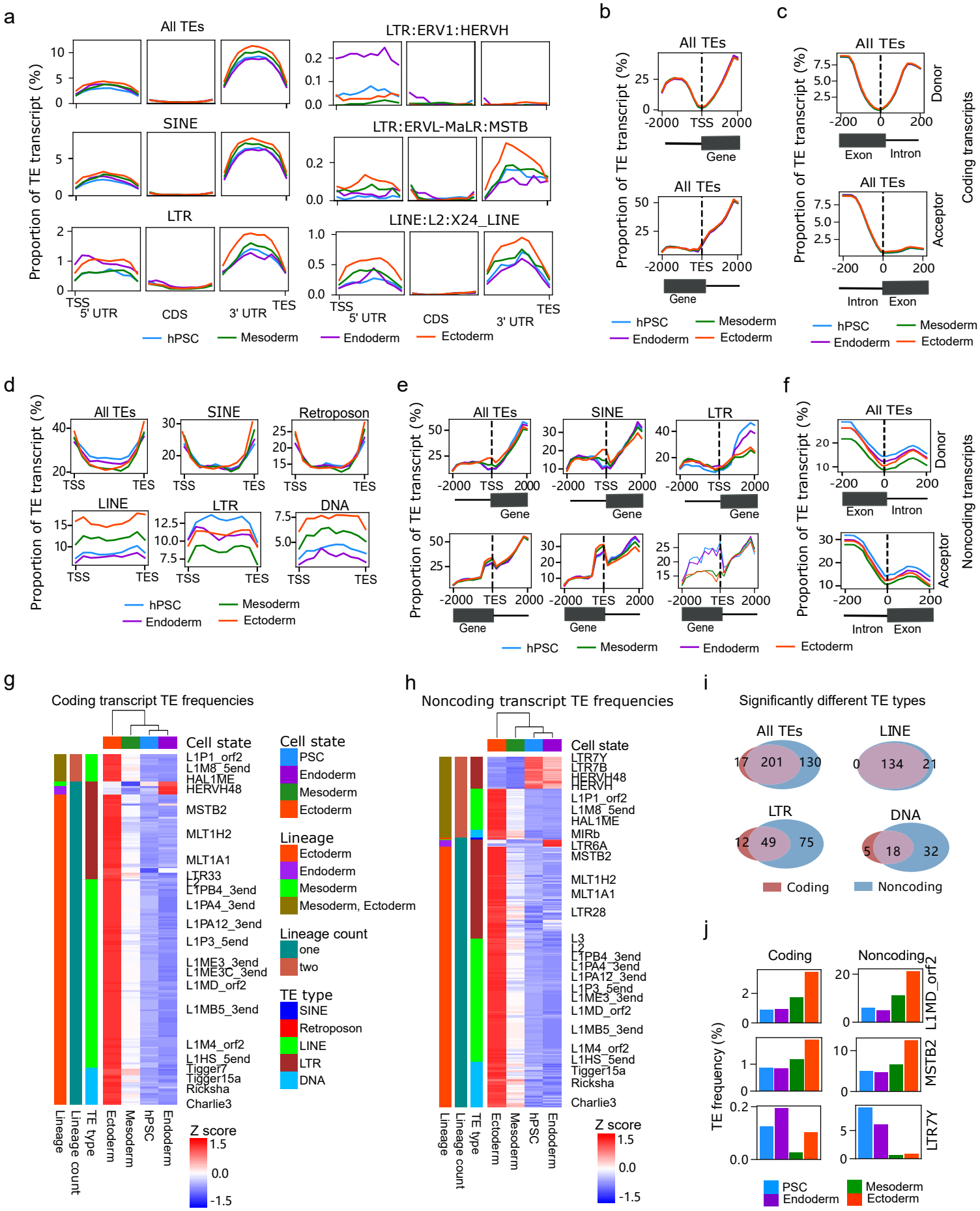
a
All TEs
LTR:ERV1:HERVH
SINE
LTR:ERVL-MaLR:MSTB
LTR
LINE:L2:X24_LINE
Proportion of TE transcript (%)
TSS
5' UTR
CDS
3' UTR
TES
hPSC
Mesoderm
Endoderm
Ectoderm
b
All TEs
Proportion of TE transcript (%)
-2000
TSS
2000
TES
Gene
c
All TEs
Donor
Exon
Intron
Acceptor
Coding transcripts
d
All TEs
SINE
Retroposon
LINE
LTR
DNA
e
All TEs
SINE
LTR
f
All TEs
Donor
Acceptor
Noncoding transcripts
g
Coding transcript TE frequencies
Cell state
L1P1_orf2
L1M8_5end
HAL1ME
HERVH48
MSTB2
MLT1H2
MLT1A1
LTR33
L2
L1PB4_3end
L1PA4_3end
L1PA12_3end
L1P3_5end
L1ME3_3end
L1ME3C_3end
L1MD_orf2
L1MB5_3end
L1M4_orf2
L1HS_5end
Tigger7
Tigger15a
Ricksha
Charlie3
Lineage
Lineage count
TE type
Ectoderm
Mesoderm
hPSC
Endoderm
Z score
1.5
0.0
-1.5
PSC
Endoderm
Mesoderm
Ectoderm
Lineage
Ectoderm
Endoderm
Mesoderm
Mesoderm, Ectoderm
Lineage count
one
two
TE type
SINE
Retroposon
LINE
LTR
DNA
h
Noncoding transcript TE frequencies
LTR7Y
LTR7B
HERVH48
HERVH
L1P1_orf2
L1M8_5end
HAL1ME
MIRb
LTR6A
MSTB2
MLT1H2
MLT1A1
LTR28
L3
L2
L1PB4_3end
L1PA4_3end
L1PA12_3end
L1P3_5end
L1ME3_3end
L1MD_orf2
L1MB5_3end
L1M4_orf2
L1HS_5end
Tigger15a
Ricksha
Charlie3
i
Significantly different TE types
All TEs
17
201
130
LINE
0
134
21
LTR
12
49
75
DNA
5
18
32
Coding
Noncoding
j
Coding
Noncoding
TE frequency (%)
L1MD_orf2
MSTB2
LTR7Y
PSC
Endoderm
Mesoderm
Ectoderm

Fig. 4

Fig. 4: TE splicing patterns and frequencies of TE-containing coding and noncoding transcripts **a**, The splicing patterns of different groups of TEs in coding transcripts. Coding transcripts were divided into 5' UTR, CDS and 3' UTR, and scaled to a uniform size. **b**, Proportion of TEs 2 kbp upstream and downstream of coding transcript TSS and TES. **c**, Proportion of TEs around 200 bp upstream and downstream of the donor and acceptor splice sites of coding transcripts. **d**, The splicing patterns of different TE types in noncoding transcripts. Each transcript is scaled to a uniform size. **e**, TE contents around the TSS and TES of noncoding transcripts. **f**, TE contents around the donor and acceptor sites in the noncoding transcripts. For panels **a-f**, each transcript was divided into 20 bins. The TE content for each bin was then computed. **g**, Heatmap of TEs with significant difference in coding transcript frequencies between hPSC and any of the three germ layers. **h**, Heatmap of TEs with significant frequency differences in the conceding transcripts of the hPSC and any of the three germ layers. For panels **g** and **h**, the statistical significance was defined as an adjusted p-value < 0.05 and odds ratio ≥ 2. For the heatmap, the z-score of the proportion of TE-containing coding and noncoding transcripts was computed. For each TE, the lineage(s) with significant difference, and the number of lineage(s) with significant differences are shown. **i**, Venn diagrams of the overlaps of the TE types with significant frequency differences in coding and noncoding transcripts. **j**, TE frequencies of selected TEs in the coding and noncoding transcripts.

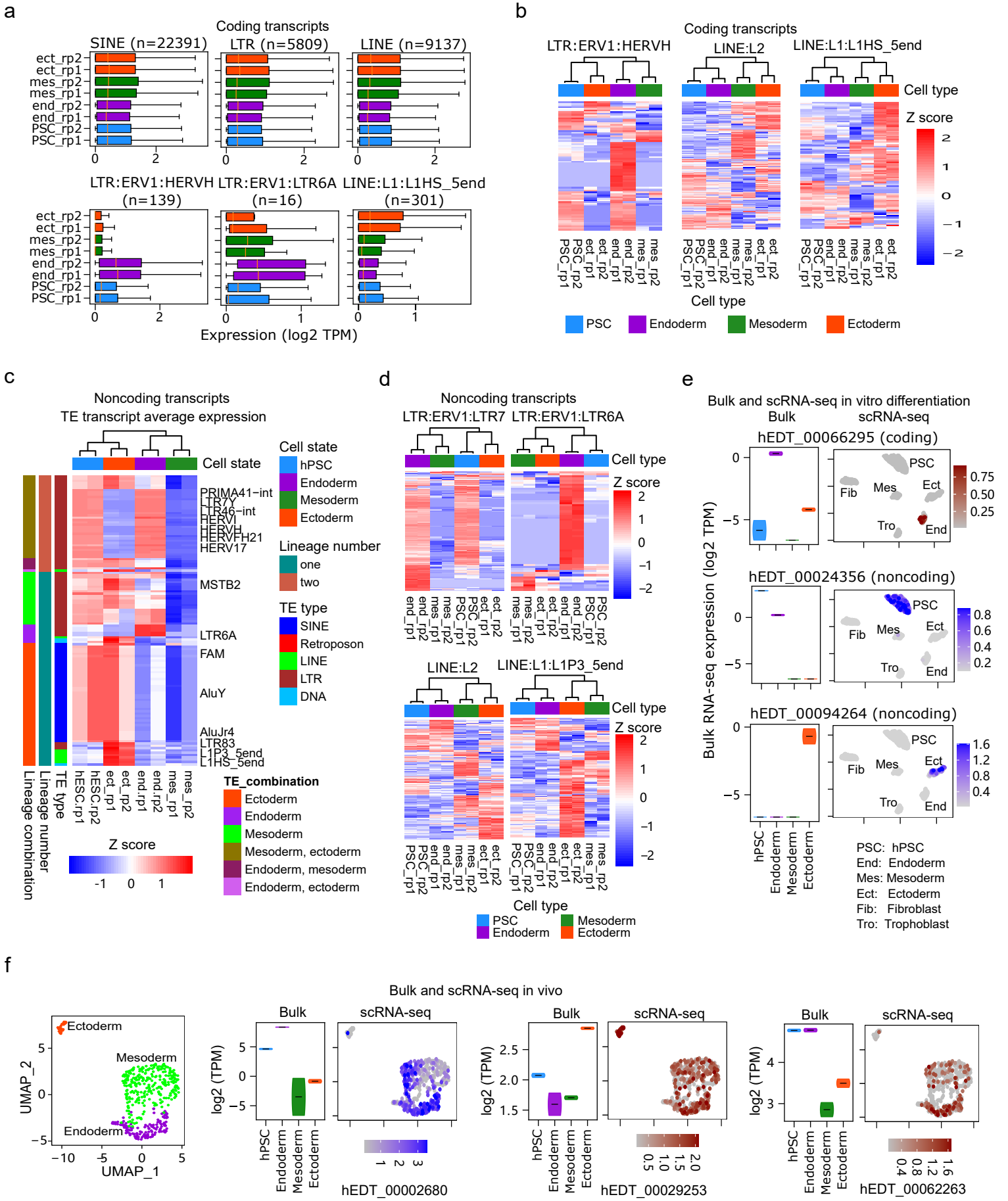
a
Coding transcripts
SINE (n=22391)
LTR (n=5809)
LINE (n=9137)
LTR:ERV1:HERVH (n=139)
LTR:ERV1:LTR6A (n=16)
LINE:L1:L1HS_5end (n=301)
ect_rp2
ect_rp1
mes_rp2
mes_rp1
end_rp2
end_rp1
PSC_rp2
PSC_rp1
Expression (log2 TPM)
b
Coding transcripts
LTR:ERV1:HERVH
LINE:L2
LINE:L1:L1HS_5end
Cell type
Z score
PSC
Endoderm
Mesoderm
Ectoderm
c
Noncoding transcripts
TE transcript average expression
Cell state
hPSC
Endoderm
Mesoderm
Ectoderm
Lineage number
one
two
TE type
SINE
Retroposon
LINE
LTR
DNA
PRIMA41−int
LTR7Y
LTR46−int
HERVI
HERVH
HERVFH21
HERV17
MSTB2
LTR6A
FAM
AluY
AluJr4
LTR83
L1P3_5end
L1HS_5end
Lineage combination
Lineage number
TE type
hESC.rp1
hESC.rp2
ect_rp1
ect_rp2
end.rp1
end.rp2
mes_rp1
mes_rp2
Z score
TE_combination
Ectoderm
Endoderm
Mesoderm
Mesoderm, ectoderm
Endoderm, mesoderm
Endoderm, ectoderm
d
Noncoding transcripts
LTR:ERV1:LTR7
LTR:ERV1:LTR6A
LINE:L2
LINE:L1:L1P3_5end
Cell type
Z score
PSC
Endoderm
Mesoderm
Ectoderm
e
Bulk and scRNA-seq in vitro differentiation
Bulk
scRNA-seq
hEDT_00066295 (coding)
hEDT_00024356 (noncoding)
hEDT_00094264 (noncoding)
Bulk RNA-seq expression (log2 TPM)
PSC
Fib
Mes
Ect
Tro
End
hPSC
Endoderm
Mesoderm
Ectoderm
PSC: hPSC
End: Endoderm
Mes: Mesoderm
Ect: Ectoderm
Fib: Fibroblast
Tro: Trophoblast
f
Bulk and scRNA-seq in vivo
Ectoderm
Mesoderm
Endoderm
UMAP_1
UMAP_2
Bulk
scRNA-seq
log2 (TPM)
hPSC
Endoderm
Mesoderm
Ectoderm
hEDT_00002680
hEDT_00029253
hEDT_00062263

Fig. 5

809 **Fig. 5: Dynamic expression patterns of TE-containing coding and noncoding transcripts**
810 **across different cell states. a**, Boxplots of the aggregate expression levels of coding transcripts
811 containing specific TEs and TE types. **b**, Heatmap showing the expression changes of
812 differentiation-induced differentially expressed coding transcripts containing the stated TEs. Each
813 row of the heatmap represents a transcript containing the indicated type of TE (top label). **c**,
814 Heatmap of aggregate changes in expression of noncoding transcripts containing different TEs.
815 Each row represents the mean of all transcripts containing a TE-type that is significantly different
816 (adjusted p-value < 0.05 and fold change >=2) between hPSC and any of the differentiated cell
817 states. For each TE, the lineage(s) with significant difference, and the number of lineage(s) with
818 significant differences are shown. **d**, Heatmap showing the change in expression levels of
819 differentiation-induced differentially expressed noncoding transcripts containing the indicated TEs.
820 Each row represents a TE-containing transcript that is differentially expressed between hPSC and
821 any of the three differentiated states. Bulk RNA-seq (left) and scRNA-seq (right) expression
822 patterns of selected TE-containing transcripts in *in vitro* (**e**) and *in vivo* (**f**) differentiation data.

823

824

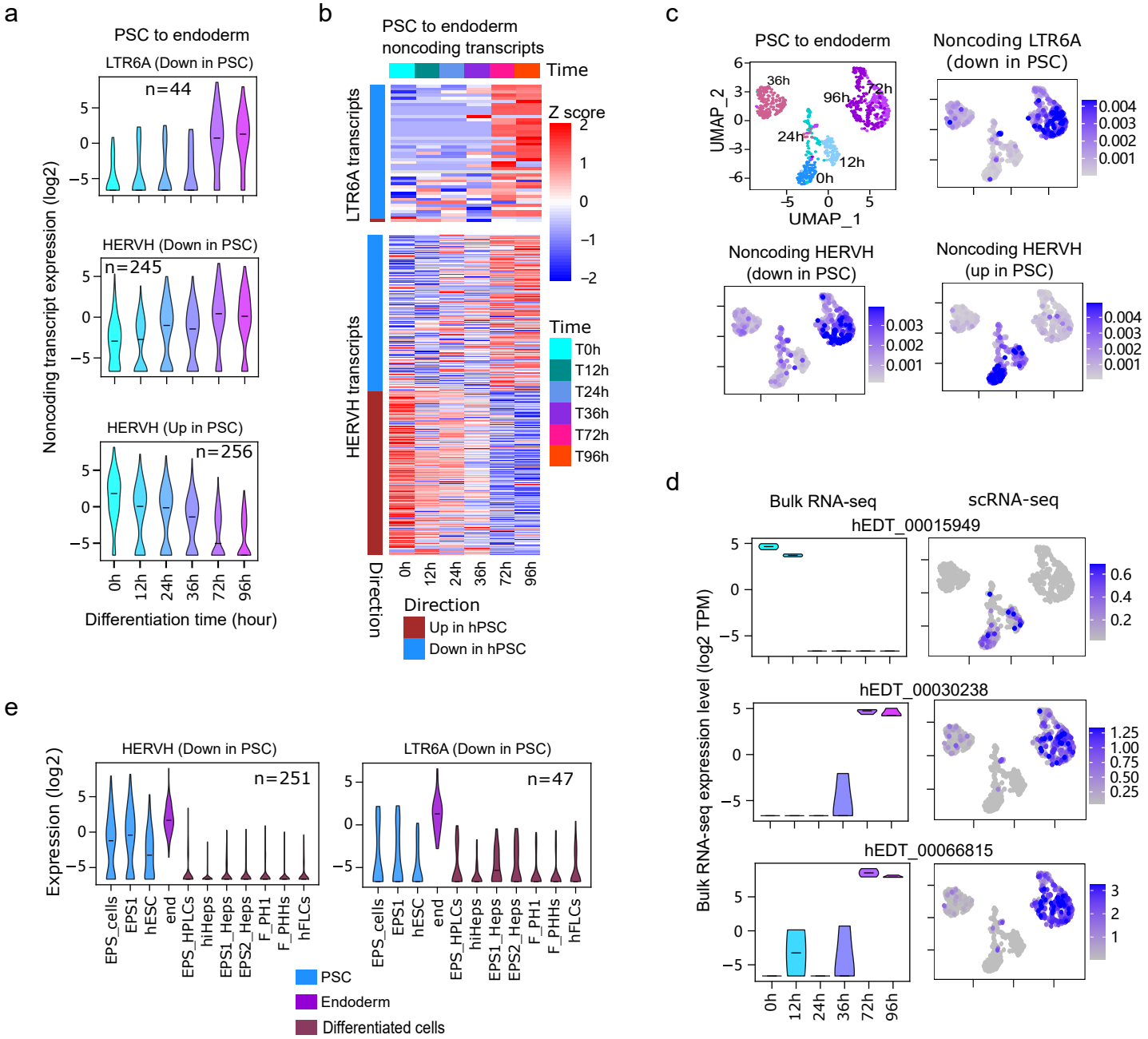
a
PSC to endoderm
LTR6A (Down in PSC)
n=44
HERVH (Down in PSC)
n=245
HERVH (Up in PSC)
n=256
Noncoding transcript expression (log2)
0h
12h
24h
36h
72h
96h
Differentiation time (hour)
b
PSC to endoderm
noncoding transcripts
Time
Z score
LTR6A transcripts
HERVH transcripts
Time
T0h
T12h
T24h
T36h
T72h
T96h
Direction
Direction
Up in hPSC
Down in hPSC
c
PSC to endoderm
UMAP_2
UMAP_1
36h
96h
72h
24h
12h
0h
Noncoding LTR6A
(down in PSC)
Noncoding HERVH
(down in PSC)
Noncoding HERVH
(up in PSC)
d
Bulk RNA-seq
scRNA-seq
hEDT_00015949
hEDT_00030238
hEDT_00066815
Bulk RNA-seq expression level (log2 TPM)
e
HERVH (Down in PSC)
n=251
LTR6A (Down in PSC)
n=47
Expression (log2)
EPS_cells
EPS1
hESC
end
EPS_HPLCs
hiHeps
EPS1_Heps
EPS2_Heps
F_PH1
F_PHHs
hFLCs
PSC
Endoderm
Differentiated cells

Fig. 6

825 **Fig. 6: Expression changes of TE-containing transcripts in endoderm lineage. a**, Violin plots
826 of bulk RNA-seq expression levels of HERVH and LTR6A-containing noncoding transcripts
827 during differentiation to endoderm. The data were from GSE75748. **b**, Heatmap showing the
828 expression dynamics of HERVH and LTR6A noncoding transcripts differentially expressed in
829 endoderm lineage. **c**, UMAP showing the single-cell expression data for a hPSC to endoderm time
830 course. Hours of differentiation are shown. The three other UMAPs show example aggregate
831 scores for transcripts containing the indicated TE types that are down-regulated in hPSCs or up-
832 regulated in hPSCs. **d**, Expression dynamics in bulk RNA-seq (left) and scRNA-seq (right) of
833 selected transcripts. **e**, Expression dynamics of endoderm-upregulated HERVH and LTR6A-
834 containing noncoding transcripts in hepatocyte-related cells, using transcripts containing the
835 indicated TE types, that are defined as down-regulated in hPSCs. (EPS: Extended pluripotent stem
836 cells; EPS1: Stage 1 EPS; EPS2: Stage 2 EPS; hESC: Human ESC (H1); HPLC: Hepatic
837 progenitor -like cells; EPS_HPLCs: EPS-derived HPLCs; hiHEPs: iPSC-derived hepatocytes;
838 EPS1_Heps: EPS1-derived hepatocytes; EPS2_Heps: EPS2-derived hepatocytes; F_PH1: Fetal
839 primary hepatocytes; F_PHHs: Fresh primary human hepatocytes; hFLC: Human fetal liver cells).

840

841

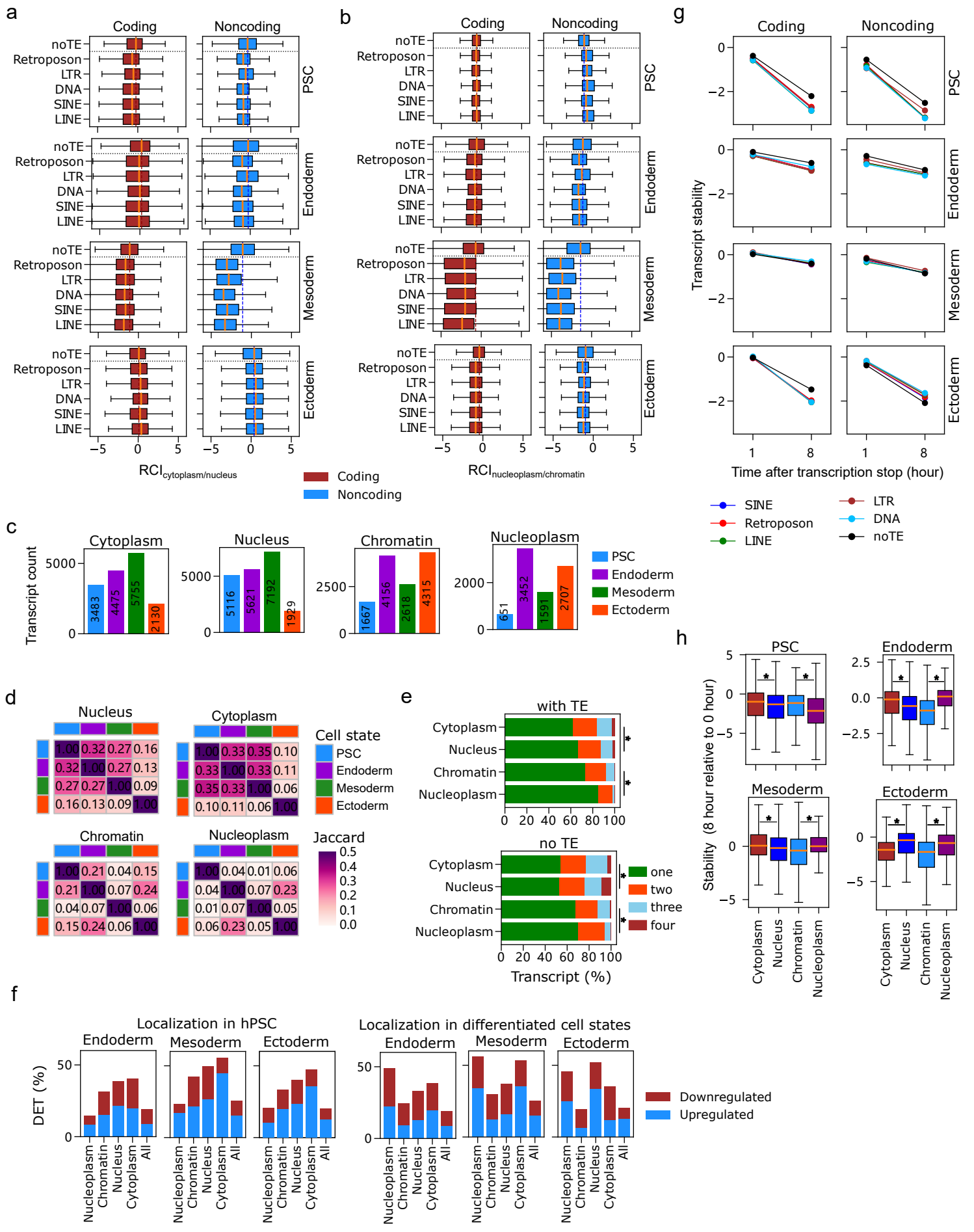
a
Coding
Noncoding
noTE
Retroposon
LTR
DNA
SINE
LINE
PSC
Endoderm
Mesoderm
Ectoderm
RCI cytoplasm/nucleus
b
RCI nucleoplasm/chromatin
Coding
Noncoding
g
Transcript stability
Time after transcription stop (hour)
SINE
Retroposon
LINE
LTR
DNA
noTE
c
Transcript count
Cytoplasm
Nucleus
Chromatin
Nucleoplasm
3483
4475
5755
2130
5116
5621
7192
1929
1667
4156
2618
4315
651
3452
1591
2707
PSC
Endoderm
Mesoderm
Ectoderm
h
Stability (8 hour relative to 0 hour)
d
Cell state
Jaccard
e
with TE
no TE
one
two
three
four
Transcript (%)
f
Localization in hPSC
Localization in differentiated cell states
DET (%)
Downregulated
Upregulated
Nucleoplasm
Chromatin
Nucleus
Cytoplasm
All

Fig. 7

842 **Fig. 7: Subcellular localization and stability of TE-containing transcripts**. **a**,
843 Cytoplasm/nucleus relative concentration index (RCI). RCI is the log2-transformed relative
844 expression in two tissues. **b**, Nucleoplasm/chromatin RCI. **c**, Bar charts showing the number of
845 transcripts localized to different subcellular structures in different cell states using individual cell
846 state assemblies. **d**, Heatmap showing subcellular localization consistency of hEDT transcripts
847 localized to different structure across cell states. **e**, Stacked bars showing the distribution of the
848 number of cell states in which transcripts are localized to a subcellular structure. (* represents Chi-
849 square significance at p value < 0.05). **f**, The proportion of localized transcripts that were
850 differentially expressed during differentiation process. The left charts show the localization in
851 hPSC while the right charts show the localization in differentiated cell states. The distributions for
852 all localizations were significantly different from the distribution of all hEDT set (Chi-square p
853 value < 0.05). **g**, Line plots showing the transcript stability of TE-containing transcripts, using time
854 from actinomycin D treated cells. The expression levels at time 1 and 8 hours, relative to the
855 expression at 0 hour are shown for each TE type across different cell states. **h**, Boxplots showing
856 the stability of transcripts that are localized to different subcellular structures across cell states (*
857 represents Mann-Whitney test significance at p value < 0.05).