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[1,3], Katerina Oleynikova[4,5], Mobolaji T. Akinwole[6], Xuemeng Zhou[1], Alexey Ruzov[4], 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]Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow, Russia.

[5]Infochemistry Scientific Center, ITMO University, St. Petersburg, Russia

[6]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 interactions.

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]. TEs have multiple evolutionary histories with different genomic distributions across multiple mammalian species[5]. Because of their repetitive nature and their underrepresentation in protein-coding genes, TEs were previously considered to have little functional importance[6,7]. In many comparative genomics studies, TE-containing genomic regions were left out and considered as genomic dark matter[8,9]. However, recent studies have shown that TEs are functionally important in diverse biological processes[2,4], including transcription[10–12], post-transcription expression regulation[13,14], transcript processing and stability[13,15,16], chromatin regulation[17,18], development[19,20] and disease progression[21,22]. While TEs have been co-opted in normal developmental processes[23,24], TEs still pose a risk to genomic integrity[4,25]. TEs are sources of mutations, can lead to genome rearrangement, interfere with normal regulatory networks and disrupt transcript coding potentials[4,26]. Therefore, TEs have both positive and negative aspects[7]. Although TEs have been studied in pluripotent stem cells[24,27] and some terminally differentiated somatic tissues[21,28], 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[29,30]. 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[31]. Thus, assembling transcripts from only short-reads is fraught with difficulty even in well-annotated species such as human[32,33], and attempting to include assembly of TE-containing transcripts is doubly difficult[31]. Previous efforts have considered global TE expression without consideration for the specific TE loci[21,34,35], nor with their transcript context. Efforts are now being made to investigate TE expression at specific loci[27,29], 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[36]. Although both short-read and long-read can be used for transcript assembly[27,30], the quality of the transcript assembly based on long-read data is superior[27,37]. 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[13]. 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[27]. 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[38]. Next, StringTie2[39,40] 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[32,41,42]. Errors might arise because of the long-read region-biased error[43], reference-induced error[44], 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[30,31]. 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[45] 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)**. Indeed, out assembly correctly identified transcripts of *POU5F1* and *SLC2A* with the necessary deepCAGE, POLR2A, polyA and epigenetic peaks (**Extended Data Figs. 1i, j**). Additionally, our assembly included previously unannotated novel transcripts such as a cluster of transcripts found on human chromosome 1 (**Fig. 1h**). We were also able to assemble new variants of known genes, such as an isoform of *LASIL* gene with a skipped exon (**Fig. 1i**). 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 hPSC *in vitro* differentiation to three germ layers using previously published protocols[46] (**Fig. 2a**). Lineage-specific marker genes were specifically up-regulated in each lineage while pluripotent markers gradually decrease along the differentiation time course (**Extended Data Fig. 2)**. To further confirm the fidelity of our differentiation experiments, we performed Fluorescence-Activated Cell Sorting (FACS) using *PAX6*, the marker for ectoderm cells. As expected, the success of the ectoderm differentiation was confirmed by the expression of *PAX6* (**Fig. 2b**). Compared to 1% in PSC, >90% of the ectoderm cells expressed *PAX6*. After the initial confirmation of our cell types, long-read sequencing was done for two technical replicates for each cell state, with each sample sequenced to 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. 2c**). These corresponded to 14,989 endoderm, 17,520 mesoderm, and 24,766 ectoderm genes (**Extended Data Fig. 3b**). The saturation analyses with down-sampling suggest that increasing the sequencing depths would lead to the discovery of more transcripts, especially the novel transcripts (**Extended Data Fig. 3c**), consistent with the previous report[47]. Since the number of assembled transcripts would be affected by the sequencing depth[31], 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 subsampled 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. 3d**). As the impact of sequencing depth tends to be more pronounced in lowly expressed transcripts[31], 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. 3e**), 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. 2d, Extended Data Fig. 3f)**. 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[27] consistently clustered with hPSC data, demonstrating the robustness of our pipelines to technical replicates and different hPSC cell lines. To further confirm the fidelity of our differentiation experiments, we retrieved 16 human iPSC, endoderm, mesoderm and ectoderm RNA-seq data sequenced on Nanopore platform[48]. We performed transcript quantification based on our merged hEDT assembly, and performed pairwise correlations between our samples and the Nanopore samples. The correlation analyses showed that the closest Nanopore sample had the same identify for all the eight PacBio samples (**Extended Data Fig. 3g)**. 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. 2 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. 3h)**. We next predicted coding and noncoding potential using FEELnc[49], 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 (v43) transcript set[50] revealed that ectoderm indeed had a lower proportion of GENCODE-known (matching) transcripts (**Extended Data Fig. 3i)**. Noncoding transcripts tended to be novel transcripts in all cell types, with a higher proportion of novel transcripts in ectoderm (**Fig. 2g**). Even at gene level, the majority of the novel genes were noncoding (**Fig. 2h**). These results highlighted the ability of assembly to identify previously unknown transcripts.

With thousands of novel transcripts, we checked for the signatures of functionality for the novel transcripts. Compared to the random sequences, both novel and matching transcripts were found to be enriched in SNPs with clinical and functional relevance (**Fig. 2g, Extended Data Fig. 3j**), however higher enrichment was found for matching transcripts. Although ectoderm enrichment was lower, the novel transcripts from all cell states had significantly higher functional SNP enrichment compared to matched random sequences (Fisher Exact p value < 0.001). We then checked if novel and matching transcripts had similar transcription factor binding motif (TFBM) enrichment in the core promoters of the transcripts. Indeed, the correlations of the TFBM enrichments between novel and matching transcripts were >0.6 for the four cell states (**Extended Data Fig. 3k**). Although the majority of the significant transcription factors (TFs) were found in both matching and novel transcripts, some TFs were found to be enriched only in novel or matching transcripts (**Extended Data Fig. 3l**), highlighting both the presence of both unique and shared transcription factors for matching and novel transcripts. These data suggest that a substantial subset of the assembled novel transcripts was functionally relevant.

To further check the reliability of the assembled transcripts, we checked the splice sites of the assembled transcripts. Across cell states, base-level conservation levels and nucleotide frequencies of the splice junctions showed that the assembled transcripts were reliable (**Extended Data Figs. 4a, b)**. The distributions of the alternative splicing events were found to be similar across the four cell states (**Extended Data Fig. 4c**). The consistency of the splicing signals and the similarity across cell states suggest that the quality of our assembled transcripts was reliable.

To investigate the overall expression dynamics of the assembled transcripts, we extracted lineage specific markers and uniformly expressed genes across the four cell states (**Fig. 2j**). The uniformly expressed genes included known housekeeping genes such as *ACTB* and *RPL5,* while the expressions of the known lineage specific markers were confirmed. As an example, we

222 confirmed that ectoderm-specific *FEZF1* had both long-read and short-read supports for ectoderm
223 specificity (**Fig. 2k**). *EOMES*, a known endoderm marker was found to be expressed only in the
224 endoderm with both long and short read support (**Extended Data Fig. 4d**). A mesoderm marker,
225 *IGFBP3*, was found to be exclusively expressed in the mesoderm (**Extended Data Fig. 4e**).
226 Interestingly, we also found some novel transcripts such as ESC.15483.4 that was highly expressed
227 in PSC with both long and short read supports (**Fig. 2l**). As expected, the unform expression of
228 *ACTB* was supported with both long and short reads. The different expression patterns could be as
229 a results of different cell states or the consequences of the differentiation factors used in the
230 experiments. To further investigate this possibility, we subjected the human embryonic skin
231 fibroblast (hESF) cells to mesoderm and endoderm differentiation conditions. Unlike in PSC, all
232 hESF-derived samples clustered together, suggesting that hESFs did not respond to the
233 differentiation process (**Extended Data Fig. 4e**). This result showed that the expression dynamics
234 of the cells actually reflected the cell states, and not just a reflection of medium treatment.

235 To investigate the heterogeneity of transcript expression at the single cell level, previously
236 published scRNA-seq data[51] containing hPSCs, and differentiated cell types was reanalyzed
237 against our hEDT assembly. UMAP clustering revealed that cells of similar states tended to cluster
238 together (**Extended Data Fig. 5a**). Meanwhile, hEDTs specifically expressed in each cell state and
239 the top 20 uniformly expressed transcripts were extracted from the long-read data (**Fig. 2j**). We
240 then checked the expression of the selected hEDTs and found that the expression patterns were
241 consistently reflected in scRNA-seq cell states (**Fig. 2m, Extended Data Figs. 5b-f)**. Both bulk
242 (long-read and short-read) and scRNA-seq data largely consistently showed that the *in vitro*
243 differentiation and bioinformatics procedures reliably captured the expressed hEDT set.

244

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

247 Having established the reliability of the *in vitro* differentiation experiments and the assembled
248 transcripts, we then investigated TE splicing patterns across the cell states. We used *nhmmer* to
249 annotate TEs inside transcript sequences based on a previously described approach[2,52]. The
250 proportions of TE-containing transcripts and RNA read in the assembled transcripts revealed that

ectoderm expressed more TE-containing transcripts (**Extended Data Fig. 6a**). As previously reported[27], most noncoding transcripts tended to contain more TEs than the coding transcripts (**Fig. 3a)**. 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 (**Fig. 3b**). 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. 3b**) even though they had fewer overall TE-containing transcripts (**Fig. 3a**), highlighting the differences across cell states.

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. 3c**). It is important to note that these TE frequencies did not reflect the genomic TE distribution (**Extended Data Fig. 6b**), suggesting the TEs are not just randomly spliced from genomic sequences. We found that the majority of SINE-containing transcripts also had retroposon motifs, suggesting that these two TE types co-exist in hEDTs (**Extended Data Fig. 6c**). Consequently, SINE-containing transcripts were considered to represent retroposon-containing transcripts in the clustering analyses. 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. 3c**). Indeed, the higher frequency of SINE and retroposon splicing is observed across cell states in both coding and noncoding transcripts. Interestingly, a number of TE-containing transcripts were expressed in a cell state-specific manner. For example, a locus containing a cluster of LTRs on human chromosome 3 is exclusively expressed in PSC (**Fig. 3d**) while another LTR-containing locus on human chromosome 17 was expressed exclusively in endoderm (**Fig. 3e**). Indeed, multiple cell state-specific TE-containing transcripts such as endoderm-specific (**Extended Data Fig. 6d**), ectoderm-

specific (**Extended Data Fig. 6e**) and mesoderm-specific (**Extended Data Fig. 6f**) were detected, highlighting cell state-specific TE expression.

Since only a small proportion of transcripts are entirely covered by TEs (**Fig. 3b**), we asked which part of the transcript TE is located. In most of the transcripts, TEs are inserted into the transcript exon; a sizable number of transcripts were TE-terminating (having TE on their TTS) while very fewer transcripts have TE-overlapping TSS (**Fig. 3f**). Across different TE types, the prevalence of exonized TE transcripts were consistent but the proportions of TE-initiating and TE-terminating were heterogenous (**Extended Data Fig. 6g**). In many cases, we found that a transcript may contain more than one TE. This is especially the case for SINE-containing transcripts with more than 50% containing multiple SINEs (**Extended Data Fig. 6h**). However, higher proportion of DNA-containing transcripts contained single TE. The analyses of the evolutionary times of inserted TEs showed that TEs of different evolutionary ages tended to be found in the same transcript (**Extended Data Fig. 6i**). This is especially so for SINEs but less so for DNAs for which more than 80% of DNA-containing transcripts had single-origin TE. Interestingly, although different TE types inserted into the transcripts seemed to have different evolutionary times, the overall patterns were similar across cell states (**Fig. 3g**). Focusing on TEs with single evolutionary origin, we found that majority of the TEs evolved in the Primates common ancestor (**Fig. 3h**). The next most common evolutionary time was in the eutherian common ancestor. Interestingly, the uniqueness of the evolutionary history of SINE-containing transcripts became more obvious when single-origin TE transcripts were considered (**Extended Data Fig. 6j**). Specifically, unlike for SINEs, the majority of the expressed LINEs, LTRs and DNAs evolved in Eutheria common ancestor, revealing the heterogeneity of TE evolutionary histories.

To investigate the consequences of TE insertion into transcripts, we focused the analyses on loci containing both TE-lacking and TE-containing transcripts. Consistent with the overall pattern, ectoderm cells tended to contain more TE-containing transcripts with TE-lacking splice variants, and SINE-containing transcripts tended to be more (**Fig. 3i**). We then attributed TE-containing transcripts to the closest TE-lacking transcripts. Interestingly, more than 75% of the TE-containing transcripts were either coding or linked to a protein-coding TE-lacking transcripts (**Fig. 3j**). LTR-containing transcripts tended to have more TE-containing noncoding transcripts for which the closest TE-lacking transcript was also noncoding. Across TE types and cell states, the

major consequence of TE insertion was "modifying" in which the TE insertion substantially affected the CDS length but did not lead to the loss of coding potential. Interestingly, we also found instances in which TE insertion led to the complete loss (disrupting) or gain (creating) of coding ability. These results suggest that TE insertion may have different consequences, especially on the coding potential, of the TE-lacking transcripts.

Because TE insertion could have serious consequences on the coding potentials, we investigated the expression dynamics of TE-containing coding and noncoding transcripts. For both coding (**Extended Data Fig. 7a**) and noncoding (**Extended Data Fig. 7b**) transcripts, SINEs and retroposons were more predominant. Whereas about 80% of noncoding transcripts in ectoderm contain a SINE element, 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. 7c**). 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. 7d**). 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. 7e**). 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[53,54]. As previously reported[55], noncoding transcripts had higher expression variability, as quantified by the coefficient of variation, across cell types (**Extended Data Fig. 7f**). Further, TE-containing coding and noncoding transcripts tended to have higher expression variability than

342 TE-free transcripts. Interestingly, the Tau coefficient which measures expression specificity
343 showed that both TE-containing coding and noncoding transcripts had higher cell-specific
344 expression than TE-free transcripts, suggesting that TE expression is regulated in a cell type-
345 specific manner. We next checked the expression variability within cell states using scRNA-seq
346 data and found that hPSC transcripts had lower expression variability than differentiated cells
347 (**Extended Data Fig. 7g**). Except for mesoderm LTR-containing transcripts, noncoding transcripts
348 tended to have higher expression variability than the coding transcripts. Notably, TE presence led
349 to higher expression variability in coding transcripts but lower variability in noncoding transcripts.
350 The differences in the expression variability might be influenced by the expression detectability,
351 especially for lowly expressed transcripts. We checked the expression detectability and found that
352 TE-containing transcripts tended to have higher detectability (**Fig. 3j**). Interestingly, while the
353 transcript detectability in TE-free noncoding transcripts was lower than that of the TE-free coding
354 transcripts, the detectability of the TE-containing coding and noncoding transcripts were more
355 similar. These data showed that the low expression variability in TE-containing noncoding
356 transcripts was not just because of the inability to detect due to the low transcript expression. Taken
357 together, these results suggest that the trajectory of TE expression is coordinated during
358 differentiation, but cell-level coding and noncoding transcript expression is dynamic.

359

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

362 Having established the dynamic expression of TE-containing transcript in different lineages, we
363 investigated the splicing patterns and insertion frequencies of different TEs in the four cell states.
364 Because of the differences in coding and noncoding transcripts, we investigated the splicing
365 patterns separately. Analysis of TE-splicing patterns in coding transcripts showed that TE
366 sequences are rare in coding sequence (CDS) regions (**Fig. 4a, Extended Data Fig. 8a**),
367 presumably reflecting the evolutionary cost of a TE inserting into and disrupting a coding sequence.
368 Across TE types and cell states, TEs were overrepresented in the 5' and 3' untranslated regions
369 (UTRs) compared to the CDS, with higher frequencies in the 3' UTRs. However, investigation of
370 specific TEs showed differences between different regions and cell states. For example, while
371 HERVH and MSTB are both LTRs, their TE splicing patterns were different, with HERVH

372 sequences being rare in the 3'UTR, whilst MSTBs were enriched (**Fig. 4a**). In addition, there was
373 a higher frequency of HERVH TE-derived sequences in endoderm cells. Interestingly, the TE
374 pattern of X24_LINE was very similar to that of MSTB, and also showed higher splicing in the
375 UTRs of ectoderm, revealing the heterogeneity of the splicing patterns of specific TEs. The
376 comparison of the MER3 and LTR6A frequencies further highlighted differences between these
377 TEs and different cell states (**Extended Data Fig. 8a**).

378 We next looked at splicing at the TSS and TES boundaries of transcripts, and interestingly,
379 while the TE splicing patterns across the TSS and TES were different, there were no substantial
380 differences across cell states (**Fig. 4b**). Across the TE types, TE proportion was low at the TSS. In
381 contrast, TEs were also reduced at the TES, although this reduction extended 5' into the transcript
382 body (**Extended Data Fig. 8b**). Similarly, TE frequencies at both the donor and acceptor splice
383 junctions were low (**Fig. 4c, Extended Data Fig. 8c**). Although there are subtle differences across
384 cell lines and TE types, the general patterns were similar (**Extended Data Fig. 8c**). Surprisingly,
385 TE frequencies of the exons at the splice sites suggest that TE insertion suppression at the exon
386 starts is stronger than the suppression at the exon ends. This supports the idea that TEs are impaired
387 from splicing into TSSs, TESs, or splice junctions directly to prevent the disruption of important
388 motifs for transcription initiation and termination and transcript splicing.

389 We next studied the properties of TEs in noncoding transcripts and found that TE splicing
390 patterns in noncoding transcripts were dynamic between TE types, cell states, and even positions
391 within the transcript (**Fig. 4d, Extended Data Fig. 8d**). Specifically, while SINE and retroposon
392 TEs are enriched towards the 5' and 3' ends of the transcripts, LINE, LTR, and DNA TEs are
393 uniformly distributed. Additionally, the overall frequencies were different across cell states (**Fig.
394 4d, Extended Data Fig. 8d**). Ectoderm noncoding transcripts were enriched for LINE and DNA
395 TEs, whilst LTRs were enriched in hPSCs, and were also high in ectoderm and endoderm (**Fig.
396 4d**). The dynamics of TE expression was more obvious for transcripts containing LTRs. For
397 example, while HERVFH21 was concentrated in the center of noncoding transcripts, endoderm
398 LTR6A had higher enrichment around 5' and 3' ends, suggesting these transcripts are noncoding
399 remnants of intact ERVs. HERVH, conversely, was uniformly distributed across the noncoding
400 transcripts in hPSC and endoderm cells (**Extended Data Fig. 8d**), highlighting the differences
401 among TEs.

402 TE frequencies around noncoding transcript TSSs were also dynamic. SINE, LTRs, and
403 retroposon were rare at the TSS (**Fig. 4e, Extended Data Fig. 8e)**. Conversely at the transcript
404 ends, SINEs and LTRs were enriched up to the TES, and then their enrichment dropped
405 substantially, suggesting that they marked transcript ends. TE sequences were absent at the TSSs
406 of noncoding transcripts, suggesting TE sequences rarely act as core promoters of TSSs. TE
407 frequencies at noncoding transcript splice junctions were also low (**Fig. 4e, Extended Data Fig.**
408 **8f)**, but not as low as in coding transcripts (**Extended Data Fig. 8f)**. Across the noncoding
409 transcript bodies, TSS, TES, and splice junctions (**Figs. 4d-f, Extended Data Figs. 8d-f**), splice
410 pattern differences based on the TE type and cell states were obvious.

411 To figure out the cell type-specific TE activities, we explored their frequencies in coding
412 and noncoding transcripts. The comparisons of the frequencies of TE-containing coding
413 transcripts identified 218 TEs that were significantly different in their frequencies upon
414 differentiation into endoderm, mesoderm, or ectoderm (**Fig. 4g**). The majority of the TEs that were
415 significantly different were enriched in ectoderm cells and were LINEs and LTRs and a few DNA
416 TEs. In contrast, some LTRs like LTR7, HERVH, HERVH48, and MTSB2 were enriched in
417 endoderm coding transcripts. On the other hand, the frequency comparison for noncoding
418 transcripts identified 331 specific TEs that were differentially enriched in at least one of the four
419 cell states (**Fig. 4h**). Similar to the coding transcripts, the majority of the differentially spliced TEs
420 in noncoding transcripts were LINEs and LTRs, and they were more frequent in the ectoderm.
421 Some LTRs such as LTR7B and HERVH were more frequent in hPSCs as previously reported[56,57]
422 and here we observed many of them in endoderm cells, suggesting these are not wholly specific
423 features of hPSCs. Indeed, LTR7B/Y/H and HERVH were enriched in endoderm cells (**Fig. 4h**),
424 supporting the idea that HERVH expression extends to cell types beyond hPSCs. HERVH
425 modulates 3-dimensional genome structure in hPSCs[58] and may also be performing this function
426 in somatic cells. Interestingly, there was a substantial overlap in the significantly enriched TE types
427 in any cell type for both coding and noncoding transcripts (**Fig. 4i**). Breaking this down by TE
428 type, the overlap is high for LINEs, with just 21-specific for noncoding transcripts and 134 LINE
429 types were enriched in both coding and noncoding. For example, although the frequencies were
430 different, the overall abundance patterns for LINE L1MD_orf2 were similar for coding and
431 noncoding transcripts (**Fig. 4j**). In contrast to LINEs, LTRs were different between coding and
432 noncoding transcripts, and only 49 LTR-types were common. This was exemplified by the LTRs

433 MSTB2 and LTR7Y which were specific to cell types **(Fig. 4j)**. Overall, these data indicate that
434 the TE content in different lineages is highly dynamic, whilst the TE sequences in coding and
435 noncoding transcripts are more similar, except for LTRs, which are cell type and transcript-type
436 specific.

437

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

440 Above, we mainly consider TEs at the family or type level. However, a significant advantage of
441 long-read sequence data is that we can place TEs into their specific transcript context. Although
442 there is a substantial overlap in the types of TEs that are found in coding and noncoding transcripts,
443 the exact genomic loci or the TE-containing transcripts are likely to be different. Potentially there
444 are two scenarios of TE expression in different cell states. The first scenario is that the same TE
445 type is contained in the same set of transcripts. The second scenario is that the same TE type is
446 expressed, but from different genomic loci, consequently producing different transcripts. To
447 explore this, we first needed to understand the relationship between TE type and expression level
448 at the individual transcript level. We used bulk short-read RNA-seq data for each lineage with the
449 long-read based hEDT reference, as the dynamic range of expression is higher for short-read data
450 than long-read data[31]. In all four cell states, the expression levels of TE-free coding transcripts
451 were significantly higher (T-test p-value < 0.05) than those of TE-containing coding transcripts
452 (**Extended Data Fig. 9a)**. TEs inserted into the 5' UTRs resulted in transcripts with the lowest
453 expression levels while 3' UTR insertion led to comparatively higher expression levels, but the
454 levels were not as high as those of the TE-free transcripts.

455 We next investigated the aggregate expression patterns and found that the overall
456 expression patterns for SINE, LTR, LINE, retroposon, and DNA coding transcripts were not
457 substantially different across the four states (**Fig. 5a**). However, differential transcript expression
458 between hPSC and the three differentiated cell states revealed that hPSC-to-endoderm
459 differentiation induced the least number of transcript changes while mesoderm differentiation
460 drove the largest coding transcript expression changes (**Extended Data Fig. 9b)**. Differentiation
461 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. 9c)**. 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. 9d, 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[31,59,60], 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. 9b**), fewer differentially expressed noncoding transcripts were induced by the differentiation process (**Extended Data Fig. 9g**), reflecting the lower 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. 9h**). 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. 9i**). 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.

492 Interestingly, several SINEs were also differentially expressed. LTR6A and a few other LTR TEs
493 were also found to be enriched for endoderm cells. The analyses of the differentially expressed
494 LTR7-containing transcripts showed that many of the transcripts are downregulated in ectoderm
495 and mesoderm cells, whilst LTR6A-containing noncoding transcripts were up in endoderm (**Fig.**
496 **5d**). Also, specific LINE-containing transcripts were specifically expressed in different lineages.
497 Interestingly, half of the TEs with significant expression differences also had significantly different
498 splicing frequencies (**Extended Data Fig. 9j**). These data support the model that similar TE types
499 are expressed from different transcripts in different cell types.

500 We next investigated the expression dynamics of TE-containing coding transcripts at the
501 single-cell level using scRNA-seq data from *in vitro* differentiated germ lineage cells[51]. The
502 scRNA-seq data recaptured the higher expression of multiple TEs in ectoderm cells and
503 downregulation of some LTRs such as LTR6A, LTR7, LTR7C, and HERVH in ectoderm and
504 mesoderm cells (**Extended Data Fig. 10a**). The scRNA-seq data also revealed that the expression
505 differences were not obvious at the level of TE types but became distinct at individual transcripts
506 (**Fig. 5e, Extended Data Fig. 10b)**. Also, single-cell analyses of the noncoding transcripts were
507 consistent with the observations from bulk data, and revealed upregulation of multiple TEs in
508 ectoderm cells (**Extended Data Figs. 10a, b)**. Several LTR TEs were downregulated in ectoderm
509 and mesoderm, while LTR6A was enriched in endoderm cells (**Fig. 5e, Extended Data Figs. 10c-**
510 **d**). As with bulk data, lineage-specific expression became more obvious when TEs were
511 considered at the transcript level (**Fig. 5e, Extended Data Fig. 10d**). These results demonstrated
512 that the observations in bulk RNA-seq were recaptured in single cells.

513 To further elaborate on the expression dynamics of TE-containing transcripts we extended
514 our analysis to scRNA-seq from *in vivo* gastrulation data[61] (**Fig. 5f, Extended Data Fig. 11a**). As
515 expected, ectoderm cells were marked by more TE-containing coding and noncoding transcripts
516 (**Extended Data Figs. 11a, b)**. Indeed, compared to other cell types, more ectoderm-expressed
517 transcripts were found in almost all of the TE types in both coding and noncoding transcripts
518 (**Extended Data Figs. 11c)**. Also, many transcripts that were identified in the *in vitro* data were
519 recaptured in the *in vivo* data (**Fig. 5d, Extended Data Fig. 11d**), indicating that *in vitro*
520 observations reflected *in vivo* phenomena. Taken together, these data revealed that whilst similar

521 TE types are expressed in all cell types, they originate from multiple lineage-specific transcripts,
522 suggesting that TEs of the same type from multiple loci are independently regulated.

523 To investigate the regulation of hEDTs, we checked the enrichment of transcription factor
524 binding motifs in the core promoter (200 bp upstream of the TSS) for the assembled transcripts
525 that were differentially expressed across different differentiation courses. We used expressed
526 transcripts without significant differentiation-induced expression changes as the control. As
527 expected for endoderm, motifs for pluripotent transcription factors (TFs) like POU5F1 (OCT4)
528 were enriched in transcripts with higher expression in PSC (down in endoderm), while the motifs
529 for endoderm marker EOMES and GATA4 were in transcripts upregulated in endoderm (**Extended**
530 **Data Fig. 12**). Some TFs like MAfF were found to be consistently enriched in transcripts that were
531 downregulated during differentiation. This data suggest that the lineages had both shared and
532 unique TFs activated during differentiation. Next, we checked the enrichments of RNA-binding
533 protein motifs in the differentially expressed transcripts using transcripts without differentiation-
534 induced expression changes as the background. However, a number of RNA-binding proteins
535 (RBPs) were enriched such as HNRNPK, PCBP2 and PCBP1 were enriched in both upregulated
536 and downregulated transcripts (**Extended Data Fig. 13**), suggesting that the RBPs might be
537 involved in cell fate determination without any particular fate preference. These analyses
538 highlighted complex transcript expression regulation during differentiation.

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

540 As the same TE types are expressed from different transcripts, we next took advantage of
541 differentiation time course data to investigate how the expressions of TE-containing transcripts
542 vary during the endoderm differentiation process. As many LTR6A and HERVH-containing
543 transcripts are specifically expressed in the endoderm or hPSCs (**Figs. 5b, d)**, we focused our
544 initial analyses on the transcripts containing these two TEs. The analyses of the time-course bulk
545 RNA-seq data showed that the activation of LTR6A noncoding transcripts happened at 72 hours
546 in the differentiation process (**Figs. 6a, b)**. HERVH-containing noncoding transcripts with higher
547 endoderm expression were upregulated as early as 24 hours but did not reach maximum expression
548 until 72 hours. Conversely, HERVH-containing noncoding transcripts with high expression in
549 hPSCs were downregulated in endoderm at as early as 12 hours and were essentially undetectable
550 at 72 hours of differentiation to endoderm. The situation was similar for HERVH-containing

coding transcripts (**Figs. 6a, Extended Data Fig. 14a**). These results demonstrate how HERVH and LTR6A can both be expressed in hPSCs and endoderm, but these TE sequences are contained 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. 14b, c**). Similar observations were found for HERVH-containing endoderm-upregulated coding transcripts (**Extended Data Fig. 14d**). Importantly, the expression of endoderm-specific HERVH-containing noncoding transcripts was not maintained in endoderm-derived terminally differentiated somatic cells (**Extended Data Fig. 15**). 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. 15**). 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

579 Previous studies have shown that TE-containing transcripts tend to localize in the nucleus[62,63]. To
580 this end, we generated RNA-seq data for subcellular fractions of hPSCs, mesoderm, endoderm,
581 and ectoderm differentiated cells for the nucleus and cytoplasm. Western blots for proteins known
582 to localize to the nucleus and cytoplasm confirmed that the intended cellular fractions were purified
583 (**Extended Data Fig. 16a**). We confirmed our previous observation that both coding and
584 noncoding TE-containing transcripts preferentially localize to the nucleus in hPSCs[27] (**Fig. 7a**).
585 This pattern was the same for mesoderm and endoderm. Surprisingly, this was not the case for
586 ectoderm, where the opposite pattern was seen: TE-containing transcripts were instead more likely
587 to localize in the cytoplasm. Indeed, analyses of TE types revealed that they all tended to
588 preferentially localize to the cytoplasm in ectoderm cells, and to the nucleus in all other cell states
589 (**Extended Data Fig. 16b**). Greater details were revealed at the transcript level. For example, LTR-
590 containing noncoding transcripts tended to have different cytoplasm/nucleus enrichment when
591 compared to other TE types in PSC and endoderm (**Extended Data Fig. 16c**).

592 Noncoding RNAs that localize to the nucleus are often recruited to chromatin[64,65]. Hence,
593 we performed RNA-seq on the nucleoplasm and chromatin fractions to understand the sub-nuclear
594 localization of TE-containing transcripts (**Extended Data Fig. 16a**). TE-containing hPSC
595 transcripts were reduced in the chromatin compartment, versus the nucleoplasm (**Fig. 7b)**. On the
596 contrary, the TE containing transcripts of endoderm, mesoderm and ectoderm tended to be more
597 enriched in the chromatin fraction (**Fig. 7b, Extended Data Fig. 16d)**. Transcript-level analyses
598 of nucleoplasm/chromatin enrichment revealed heterogeneity across coding potentials, cell states
599 and TE types (**Extended Data Fig. 16e)**. These data reveal that the sub-cellular and sub-nuclear
600 localization of TE-containing transcripts is cell type-specific, and retention of TE-containing
601 transcripts in the nucleus is not a general feature.

602 To further investigate the subcellular transcript localization, we compared the expression
603 levels of different transcripts to obtain transcripts enriched in specific sub-cellular fractions (see
604 Supplementary methods). Pairwise comparisons of nucleus-cytoplasm and chromatin-
605 nucleoplasm expression patterns identified transcripts that were enriched in specific sub-cellular
606 fractions (**Extended Data Fig. 17)**. The number of transcripts localized to the sub-cellular
607 compartments varied across cell types (**Fig. 7c**). Importantly, compared to the differentiated cells,
608 hPSCs had fewer transcripts differentially localized to the nucleoplasm and chromatin. The

609 distribution of coding and noncoding transcripts revealed both cell state and location-specific
610 differences. Specifically, while nucleoplasm-localized transcripts had the highest noncoding
611 proportion in hPSC, chromatin-localized transcripts had the highest noncoding proportion in
612 differentiated cells (**Extended Data Fig. 18a**). Similarly, the distribution of TE-containing
613 transcripts varied across cell types, subcellular location, and protein-coding ability (**Extended**
614 **Data Fig. 18b**). This effect was also transcript-specific, as few transcripts overlapped in the distinct
615 sub-cellular compartments between the different cell types (**Extended Data Fig. 19**). Indeed, using
616 a Jaccard similarity measure, transcripts that localized to the nucleus or cytoplasm were more
617 consistent in hPSC, endoderm and mesoderm but more divergent in ectoderm (**Figs. 7d**).
618 Interestingly, we found that the distribution of transcripts based on the number of cells in which
619 they were localized varied across subcellular structures (**Fig. 7e**). The most unique chromatin
620 localization was found in the mesoderm (**Figs. 7d**). Overall, nucleoplasm localization was the least
621 consistent across the cell states. To further explore the relationship between transcript subcellular
622 localization and cell state conversion, we checked the proportion of differentiation-induced DETs
623 among the transcripts localized to various subcellular structures across cell states and found that
624 subcellular localization significantly influenced the proportion of DETs (**Fig. 7f**), as the
625 distribution of DETs were significantly different in localized transcripts compared to overall
626 transcripts (Chi-square p-value < 0.05). For hPSCs, nucleoplasm-enriched transcripts had the
627 lowest proportion of DETs. For differentiated cells, however, the lowest proportion of DETs was
628 found in chromatin-localized transcripts. These data indicate that the localization of TE-containing
629 transcripts is cell type-specific. Overall, compared to the TE-free transcripts, we found that TE-
630 containing transcripts were more cell type-specific (**Fig. 7e**), with hPSCs having fewer chromatin-
631 enriched TE transcripts.

632 An important question was how the localization of the transcript to different subcellular
633 structure was established. We hypothesized that the activities of RBPs might contribute to the
634 transcript localization. Consequently, we checked the RBP enrichment of localized transcripts with
635 expressed transcripts with no significant bias for each cell state. Interestingly, we found multiple
636 RBPs that were enriched in transcripts localized to specific subcellular structures (**Extended Data**
637 **Fig. 20**). Interestingly, RBPs that were associated with chromatin localization such as SAMD4A,
638 PCBP2 and LIN28A were also found to be associated with nucleus localization across different
639 cell states. Also, some RBPs like PABPN1 were found to be associated with cytoplasm and

640 nucleoplasm localization across cell states. These results suggest that chromatin enrichment might
641 contribute to nucleus enrichment, and that RBPs might be implicated in transcript localization to
642 different subcellular structures.

643

644 **TE presence influences transcript stability**

645 As subcellular localization can influence degradation, we wondered if this would impact transcript
646 half-life in a TE, sub-cellular localization, and cell state-dependent manner. We performed RNA-
647 seq in differentiated cells treated with Actinomycin D to block transcription for 1 and 8 hours and
648 measured the transcript levels relative to the 0-hour time-point. Surprisingly, there were divergent
649 patterns of transcript half-life across TE types, coding potentials, and cell states (**Fig. 7g**). The
650 lowest stability was found in hPSCs and endoderm. Interestingly, the impact of TE presence was
651 also dynamic. While the TE-free transcripts were more stable in hPSC, ectoderm, and endoderm,
652 this was not the case in mesoderm (**Fig. 7g, Extended Data Fig. 21a, b**). TE-containing ectoderm
653 transcripts also contrasted with hPSCs and endoderm, as whilst TE-containing coding transcripts
654 were less stable, TE-containing noncoding transcripts were more stable than TE-free transcripts
655 (**Fig. 7g**, **Extended Data Fig. 21a, b**). These revealed that transcript stability varied across cell
656 states and transcript coding ability.

657 Finally, we checked if TE presence had any substantial influence on the stability of
658 transcripts with differential subcellular localization. The stability at 8 hours after transcription stop
659 showed differences across transcripts localized to different subcellular structures (**Fig. 7h**). While
660 nuclear-localized transcripts were more stable than cytoplasm-localized transcripts in the ectoderm,
661 the opposite was true in the other cell states. Conversely, chromatin-localized transcripts were
662 more stable than nucleoplasm-localized transcripts in hPSCs, but the nucleoplasm-localized
663 transcripts were more stable than chromatin-localized transcripts in differentiated cells.
664 Interestingly, some differences in stability can be observed as early as 1 hour after transcription
665 stop (**Extended Data Fig. 21c**). Moreover, the comparison of stabilities of TE-free transcripts to
666 those containing TEs identified multiple TEs associated with significant differences in stability
667 after 8 hours (**Extended Data Fig. 21d**). In hPSCs, TE presence mostly led to lower stability. In
668 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[27,36]. Here, we used long-read RNA-seq technology to generate an accurate TE-containing transcript assembly[30,31]. 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[10,11,66,67]. 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[34,68–70]. 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[11,57]. 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 [27,71,72], we found that TE presence leads to lower expression levels, probably due to the activities of RNA-binding proteins[16,72]. Indeed, SINE elements are overrepresented in 3' UTRs, and STAU1 promotes mRNA decay by targeting those Alu elements[16]. 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. Whether all the TE expression changes are drivers or footprints of cell differentiation require further detailed studies. However, a number of studies have demonstrated the roles of TE-containing transcripts in pluripotency[73–75], cardiomyocyte development[76] and neurogenesis[77–79]. 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.F. 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

1. Hutchins, A. P. & Pei, D. Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. *Sci Bull (Beijing)* **60**, 1722–1733 (2015).
2. Ma, G., Babarinde, I. A., Zhou, X. & Hutchins, A. P. Transposable Elements in Pluripotent Stem Cells and Human Disease. *Front Genet* (2022).
3. Biémont, C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **186**, 1085–93 (2010).
4. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 1–12 (2018).
5. Osmanski, A. B. *et al.* Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science (1979)* **380**, (2023).
6. Kim, Y.-J., Lee, J. & Han, K. Transposable Elements: No More 'Junk DNA'. *Genomics Inform* **10**, 226–33 (2012).
7. Naville, M. *et al.* Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clinical Microbiology and Infection* **22**, 312–323 (2016).

754 8. Kim, Y.-J., Lee, J. & Han, K. Transposable Elements: No More 'Junk DNA'. *Genomics Inform* **10**, 226–
755 33 (2012).

756 9. Babarinde, I. A. & Saitou, N. Heterogeneous Tempo and Mode of Conserved Noncoding Sequence
757 Evolution among Four Mammalian Orders. *Genome Biol Evol* **5**, 2330–2343 (2013).

758 10. Wang, J., Huang, J. & Shi, G. Retrotransposons in pluripotent stem cells. *Cell Regeneration* **9**,
759 (2020).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

836 47. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**,
837 2184–2185 (2012).

838 48. Dobner, J., Diecke, S., Krutmann, J., Prigione, A. & Rossi, A. Reassessment of marker genes in
839 human induced pluripotent stem cells for enhanced quality control. *Nat Commun* **15**, 8547
840 (2024).

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

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

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

847 52. Wheeler, T. J. & Eddy, S. R. Nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**,
848 2487–2489 (2013).

849 53. Sakashita, A. *et al.* Transcription of MERVL retrotransposons is required for preimplantation
850 embryo development. *Nat Genet* **55**, (2023).

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

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

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

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

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

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

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

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

867 62. Carlevaro-Fita, J. *et al.* Ancient exapted transposable elements promote nuclear enrichment of
868 human long noncoding RNAs. *Genome Res* **29**, (2019).

869 63. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs
870 in human cells. *Nature* **555**, 107–111 (2018).

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

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

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

876 67. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev* **28**,
877 (2014).

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

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

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

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

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

888 73. Sexton, C. E., Tillett, R. L. & Han, M. V. The essential but enigmatic regulatory role of HERVH in
889 pluripotency. *Trends in Genetics* vol. 38 Preprint at https://doi.org/10.1016/j.tig.2021.07.007
890 (2022).

74. 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).

75. Göke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, (2015).

76. Zhang, R. *et al.* A primate-specific endogenous retroviral envelope protein sequesters SFRP2 to regulate human cardiomyocyte development. *Cell Stem Cell* (2024) doi:10.1016/j.stem.2024.07.006.

77. Sekine, K., Onoguchi, M. & Hamada, M. Transposons contribute to the acquisition of cell type-specific cis-elements in the brain. *Commun Biol* **6**, (2023).

78. Mustafin, R. N. & Khusnutdinova, E. K. Involvement of transposable elements in neurogenesis. *Vavilovskii Zhurnal Genetiki i Selektsii* vol. 24 Preprint at https://doi.org/10.18699/VJ20.613 (2020).

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

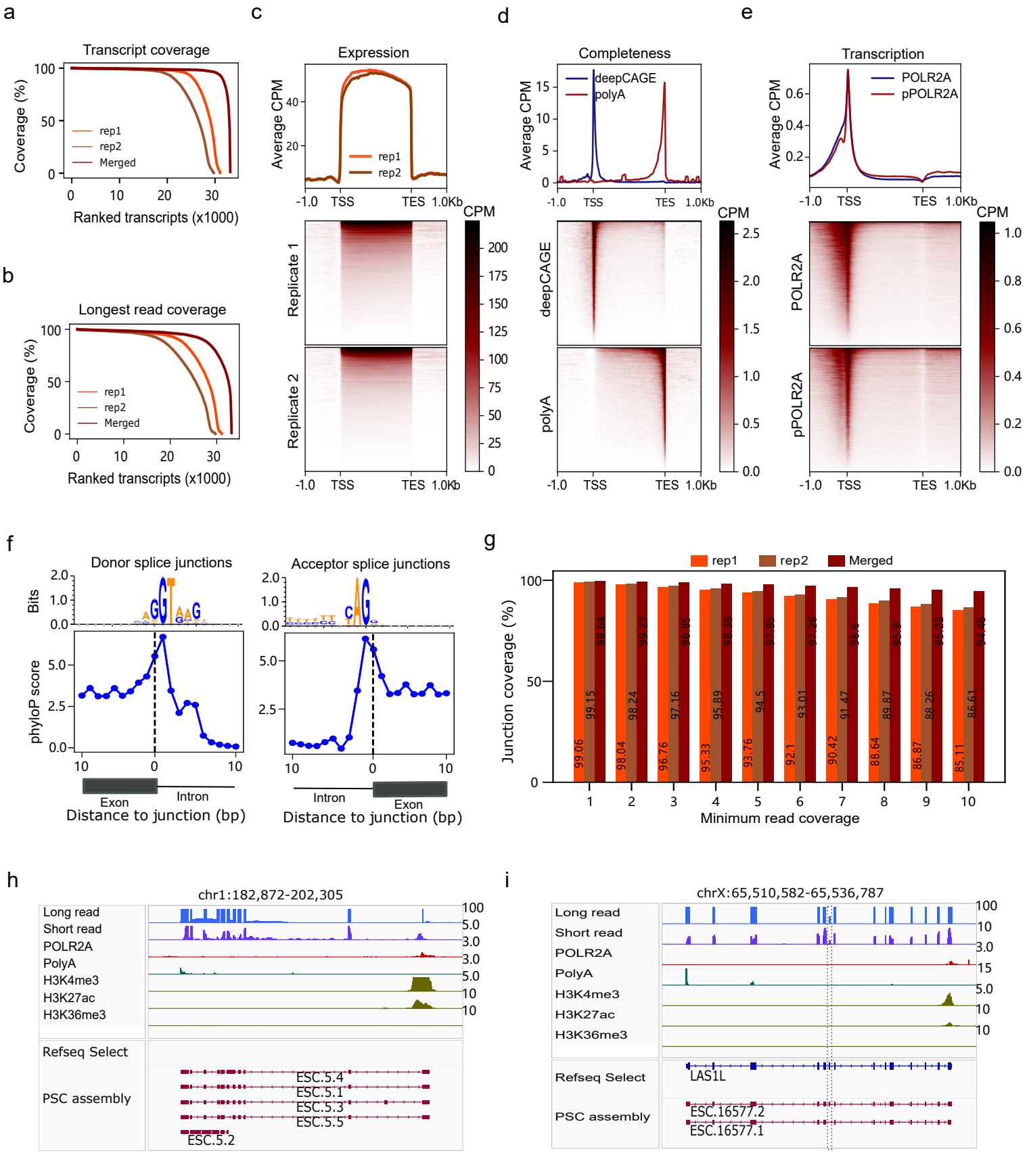
a
Transcript coverage
Coverage (%)
rep1
rep2
Merged
Ranked transcripts (x1000)
b
Longest read coverage
Coverage (%)
rep1
rep2
Merged
Ranked transcripts (x1000)
c
Expression
Average CPM
rep1
rep2
Replicate 1
Replicate 2
CPM
-1.0 TSS TES 1.0Kb
d
Completeness
Average CPM
deepCAGE
polyA
CPM
e
Transcription
Average CPM
POLR2A
pPOLR2A
CPM
f
Donor splice junctions
Acceptor splice junctions
Bits
phyloP score
Exon
Intron
Distance to junction (bp)
g
rep1
rep2
Merged
Junction coverage (%)
Minimum read coverage
h
chr1:182,872-202,305
Long read
Short read
POLR2A
PolyA
H3K4me3
H3K27ac
H3K36me3
Refseq Select
PSC assembly
ESC.5.4
ESC.5.1
ESC.5.3
ESC.5.5
ESC.5.2
i
chrX:65,510,582-65,536,787
Long read
Short read
POLR2A
PolyA
H3K4me3
H3K27ac
H3K36me3
Refseq Select
LAS1L
PSC assembly
ESC.16577.2
ESC.16577.1

Fig. 1

908 **Fig. 1: Quality assessment of the assembled hPSC long-read transcriptome**. **a**, Consensus read
909 coverage of the assembled PSC transcripts using all consensus reads. **b**, The longest consensus
910 read coverage for the assembled PSC transcripts. **c**, Pileups for the consensus long reads for the
911 assembled hPSC transcripts across the exons from the TSS to the TES. Each transcript is scaled to
912 a uniform size. **d**, DeepCAGE (5' end) and polyA tail (3' end) pileups of the assembled transcripts,
913 from the TSS to the TES scaled to a uniform size and not including intronic regions. PolyA data
914 were from GSE138759 and GSE111134 while deepCAGE data were from GSE34448 and
915 GSE61264. **e**, Density pileup of POLR2A and phosphorylated POLR2A (pPOLR2A) for the
916 assembled hPSC transcripts. The data were from GSE242645. **f**, Position weight matrix nucleotide
917 frequencies and phyloP conservation scores of the 10-bp positions around the donor and acceptor
918 splice junctions of the assembled multi-exon hPSC transcripts **g**, Proportion of hPSC junctions
919 covered by indicated minimum number of reads. Integrative genomics views showing multiple
920 novel transcripts in a locus (**h**) and a variant transcript made by a skipped exon of LASIL gene (**i**).
921 The position of a skipped exon is marked by dotted line in **Fig. 1i**.

922

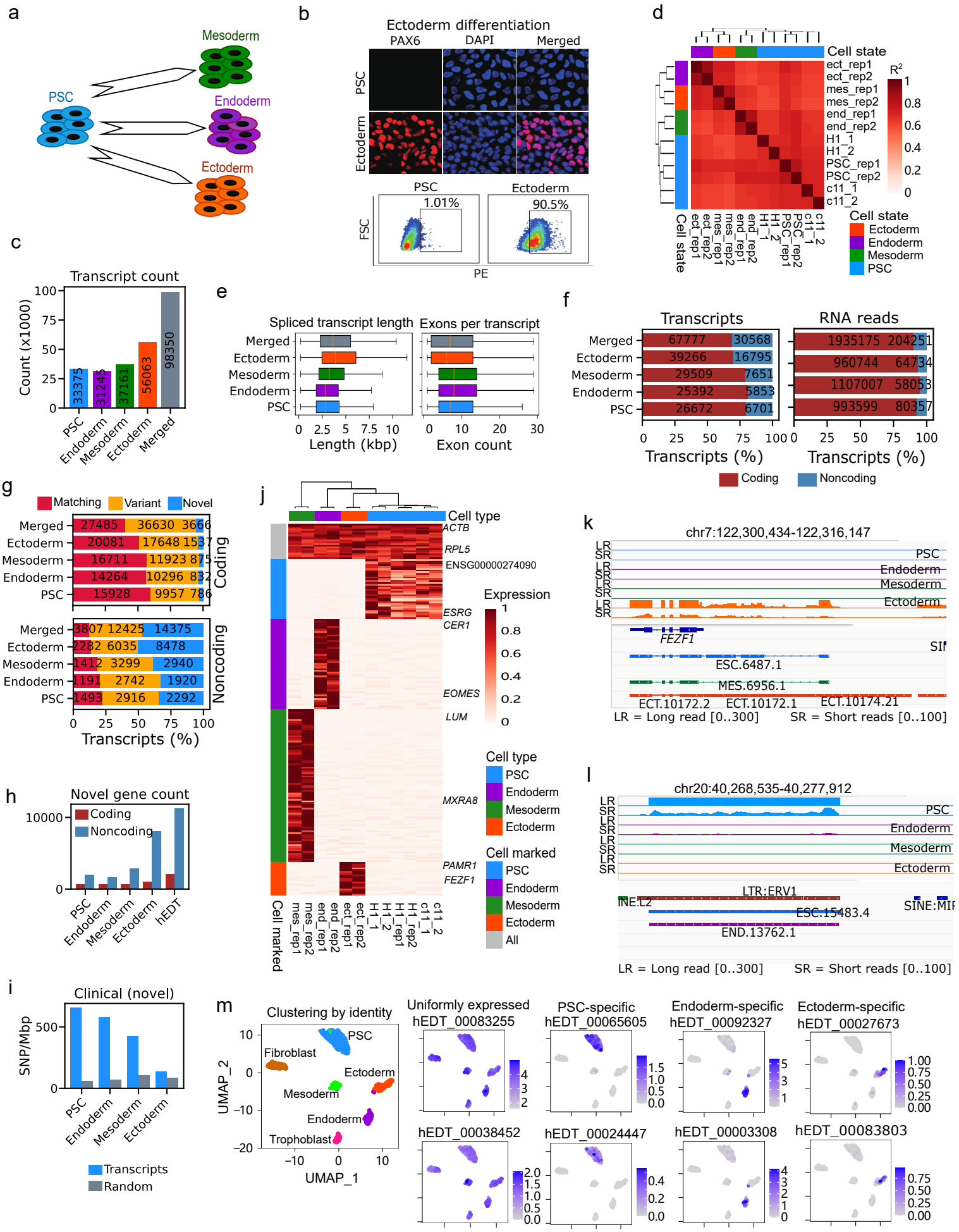
a
Mesoderm
PSC
Endoderm
Ectoderm
b
Ectoderm differentiation
PAX6
DAPI
Merged
PSC
Ectoderm
PSC
1.01%
Ectoderm
90.5%
FSC
PE
d
Cell state
ect_rep1
ect_rep2
mes_rep1
mes_rep2
end_rep1
end_rep2
H1_1
H1_2
PSC_rep1
PSC_rep2
c11_1
c11_2
R²
Cell state
Ectoderm
Endoderm
Mesoderm
PSC
c
Transcript count
Count (x1000)
33375
31245
37161
56063
98350
PSC
Endoderm
Mesoderm
Ectoderm
Merged
e
Spliced transcript length
Exons per transcript
Merged
Ectoderm
Mesoderm
Endoderm
PSC
Length (kbp)
Exon count
f
Transcripts
RNA reads
Merged 67777 30568 1935175 204251
Ectoderm 39266 16795 960744 64734
Mesoderm 29509 7651 1107007 58053
Endoderm 25392 5853 993599 80357
PSC 26672 6701
Transcripts (%)
Coding
Noncoding
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
Transcripts (%)
h
Novel gene count
Coding
Noncoding
PSC
Endoderm
Mesoderm
Ectoderm
hEDT
j
Cell type
ACTB
RPL5
ENSG00000274090
ESRG
CER1
EOMES
LUM
MXRA8
PAMR1
FEZF1
Expression
Cell type
PSC
Endoderm
Mesoderm
Ectoderm
Cell marked
PSC
Endoderm
Mesoderm
Ectoderm
All
Cell marked
k
chr7:122,300,434-122,316,147
LR
SR
PSC
Endoderm
Mesoderm
Ectoderm
FEZF1
ESC.6487.1
MES.6956.1
ECT.10172.2
ECT.10172.1
ECT.10174.21
LR = Long read [0..300]
SR = Short reads [0..100]
l
chr20:40,268,535-40,277,912
LR
SR
PSC
Endoderm
Mesoderm
Ectoderm
LTR:ERV1
INE:L2
ESC.15483.4
END.13762.1
LR = Long read [0..300]
SR = Short reads [0..100]
i
Clinical (novel)
SNP/Mbp
PSC
Endoderm
Mesoderm
Ectoderm
Transcripts
Random
m
Clustering by identity
PSC
Fibroblast
Ectoderm
Mesoderm
Endoderm
Trophoblast
UMAP_1
UMAP_2
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

Fig. 2: Expression dynamics of the assembled human early development transcripts. **a**, Simplified experimental design for the *in vitro* differentiation of human hPSC to each of the three embryonic germ layers. The detailed experimental procedures are in the Supplementary Methods. **b**, Fluorescence-activated cell sorting results using PAX6 to demonstrate the successful differentiation of PSC to ectoderm. **c**, The numbers of transcripts assembled from each cell state and the merged transcript assembly. **d**, The expression correlation heatmap of the samples made from long-read quantification. H1_1, H1_2, c11_1 and c11_2 samples were from previously reported data[27]. **e**, Boxplots showing the spliced transcript lengths (left) and number of exons per assembled transcript (right) for the assembled transcripts. Boxplots show the mean (red central bar), and second and third quartiles, and the whiskers show 1.5 times the interquartile ranges. Kruskal-Wallis p value = $2.37 \times 10^{-28}$for spliced transcript lengths and Kruskal-Wallis p value = 1.86×10^{-173} exon per transcript. **f**, Stacked bar chart showing the proportions and numbers of coding and noncoding transcripts in the hEDT assemblies. The left plot shows the distribution of the assembled transcript counts while the right plot shows the distribution of the total RNA read counts from the long-read quantifications. **g**, Stacked bar chart showing the proportions and numbers of assembled transcripts based on the similarity to version 43 of the GENCODE assembly. Matching transcript completely matches to a known transcript (including all exons and splice junctions), while a variant transcript partially overlaps a known GENCODE transcript. Novel transcript does not overlap any known GENCODE transcript exon. **h**, The number of novel coding and noncoding genes identified in each cell state. Because a gene might have more than one transcript, a gene could both be coding and noncoding. **i**, The enrichment of SNPs with clinical relevance in novel transcripts. The enrichment for matched background sequences (with equal lengths and exon structures are also shown). In all cell states, novel transcripts had significantly higher enrichment of clinical SNPs (Fisher Exact p value < 0.001). **j**, Heatmap showing the uniformly expressed transcripts and transcripts specific to each cell state based on the long-read quantification. For each transcript, the expression was normalized by the highest expression level. Integrative genomics views showing the expression of ectoderm enriched transcripts (**k**) and PSC-enriched novel transcript (**l**). **m**, Single-cell RNA-seq expression showing uniformly expressed transcripts and marker transcripts specific to the indicated cell states.

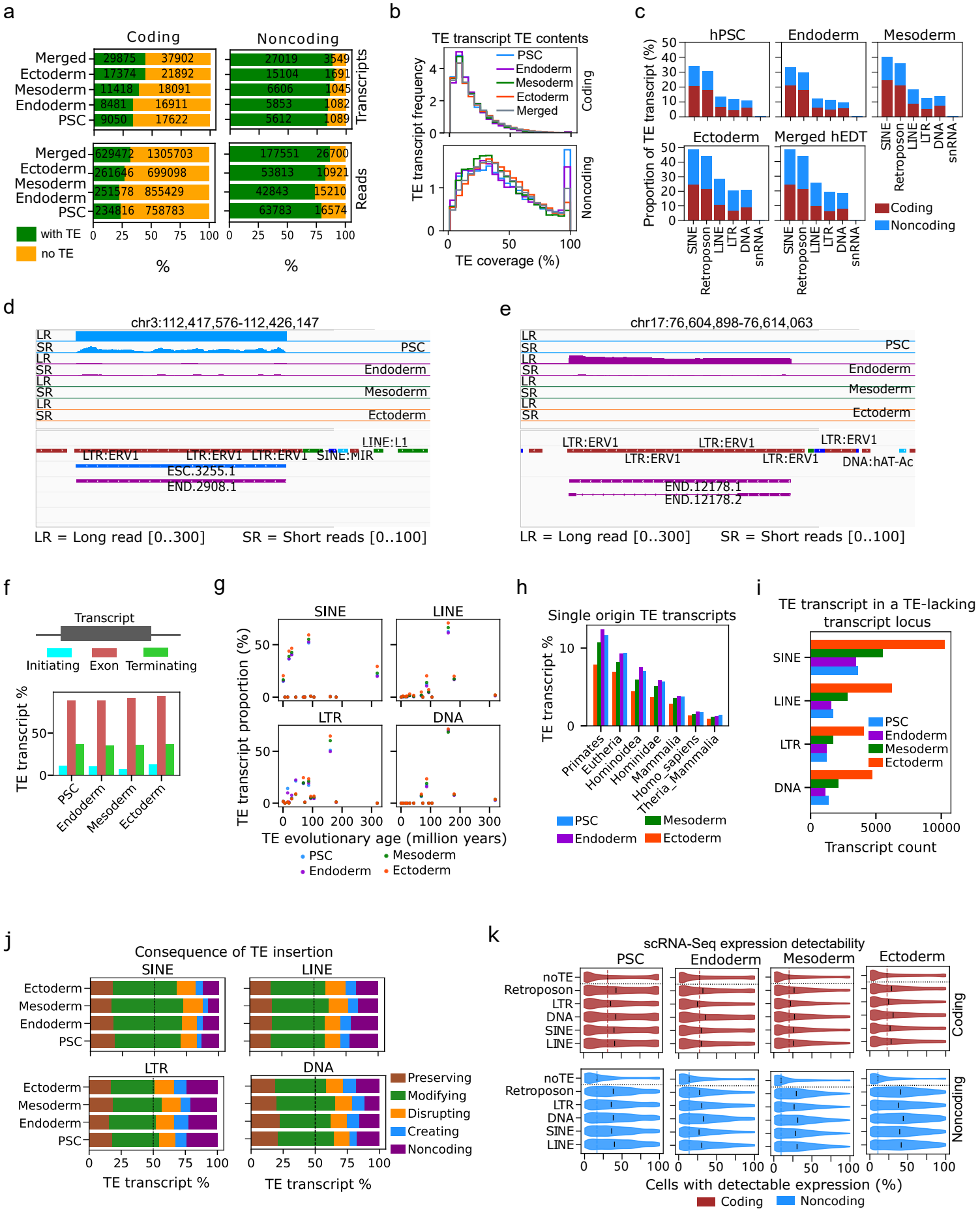
a
Coding
Noncoding
Merged
Ectoderm
Mesoderm
Endoderm
PSC
29875 37902
17374 21892
11418 18091
8481 16911
9050 17622
27019 3549
15104 1691
6606 1045
5853 1082
5612 1089
Transcripts
629472 1305703
261646 699098
251578 855429
234816 758783
177551 26700
53813 10921
42843 15210
63783 16574
Reads
with TE
no TE
%
b
TE transcript TE contents
PSC
Endoderm
Mesoderm
Ectoderm
Merged
Coding
Noncoding
TE transcript frequency
TE coverage (%)
c
hPSC
Endoderm
Mesoderm
Ectoderm
Merged hEDT
Proportion of TE transcript (%)
SINE
Retroposon
LINE
LTR
DNA
snRNA
Coding
Noncoding
d
chr3:112,417,576-112,426,147
LR
SR
PSC
Endoderm
Mesoderm
Ectoderm
LINE:L1
LTR:ERV1
SINE:MIR
ESC.3255.1
END.2908.1
LR = Long read [0..300]
SR = Short reads [0..100]
e
chr17:76,604,898-76,614,063
LTR:ERV1
DNA:hAT-Ac
END.12178.1
END.12178.2
LR = Long read [0..300]
SR = Short reads [0..100]
f
Transcript
Initiating
Exon
Terminating
TE transcript %
PSC
Endoderm
Mesoderm
Ectoderm
g
SINE
LINE
LTR
DNA
TE transcript proportion (%)
TE evolutionary age (million years)
PSC
Endoderm
Mesoderm
Ectoderm
h
Single origin TE transcripts
TE transcript %
Primates
Eutheria
Hominoidea
Hominidae
Mammalia
Homo_sapiens
Theria_Mammalia
PSC
Endoderm
Mesoderm
Ectoderm
i
TE transcript in a TE-lacking transcript locus
SINE
LINE
LTR
DNA
PSC
Endoderm
Mesoderm
Ectoderm
Transcript count
j
Consequence of TE insertion
SINE
LINE
LTR
DNA
Ectoderm
Mesoderm
Endoderm
PSC
Preserving
Modifying
Disrupting
Creating
Noncoding
TE transcript %
k
scRNA-Seq expression detectability
PSC
Endoderm
Mesoderm
Ectoderm
noTE
Retroposon
LTR
DNA
SINE
LINE
Coding
Noncoding
Cells with detectable expression (%)
Coding
Noncoding

Fig. 3

954 **Fig. 3: Divergent TE expression dynamics across coding and noncoding transcripts**. **a**,
955 Stacked bar charts showing the proportion and number of TE-containing transcripts in the coding
956 (left) and noncoding (right) transcripts in the hEDT assemblies. The upper plots show the
957 assembled transcript counts while the lower plots show the RNA read counts from long-read
958 transcript quantification. **b**, Histograms showing the distributions of TE coverages (percent of
959 sequence that consists of TE) for both coding and noncoding transcripts. **c**, Stacked bar charts
960 showing the proportion of TE-containing transcripts for coding and noncoding transcripts in the
961 assembled hEDTs. Integrative genomics views showing different loci of PSC-enriched (**d**) and
962 endoderm-enriched (**e**) LTR-containing novel transcripts. **f**, Bar plots showing the proportion of
963 TE-initiating, -exonized and -terminating transcripts. g, Scatter plots showing the proportion of TE
964 transcripts that evolved at different evolutionary timepoints. The evolutionary ages were estimated
965 based on the phylogeny of the species for the TE. **h**, Proportion of TE-transcripts of single age
966 category. **i**, The number of TE-containing transcripts with at least one TE-lacking transcript of the
967 same gene. **j**, The consequences of TE insertion based on the comparison between TE-containing
968 and the closest TE-lacking transcript of the same gene. **k**, Single-cell RNA-seq expression
969 detectability of different groups of transcripts within the cell states. The red and blue dotted lines
970 represent the median values for the coding and noncoding transcripts without TEs.

971

972

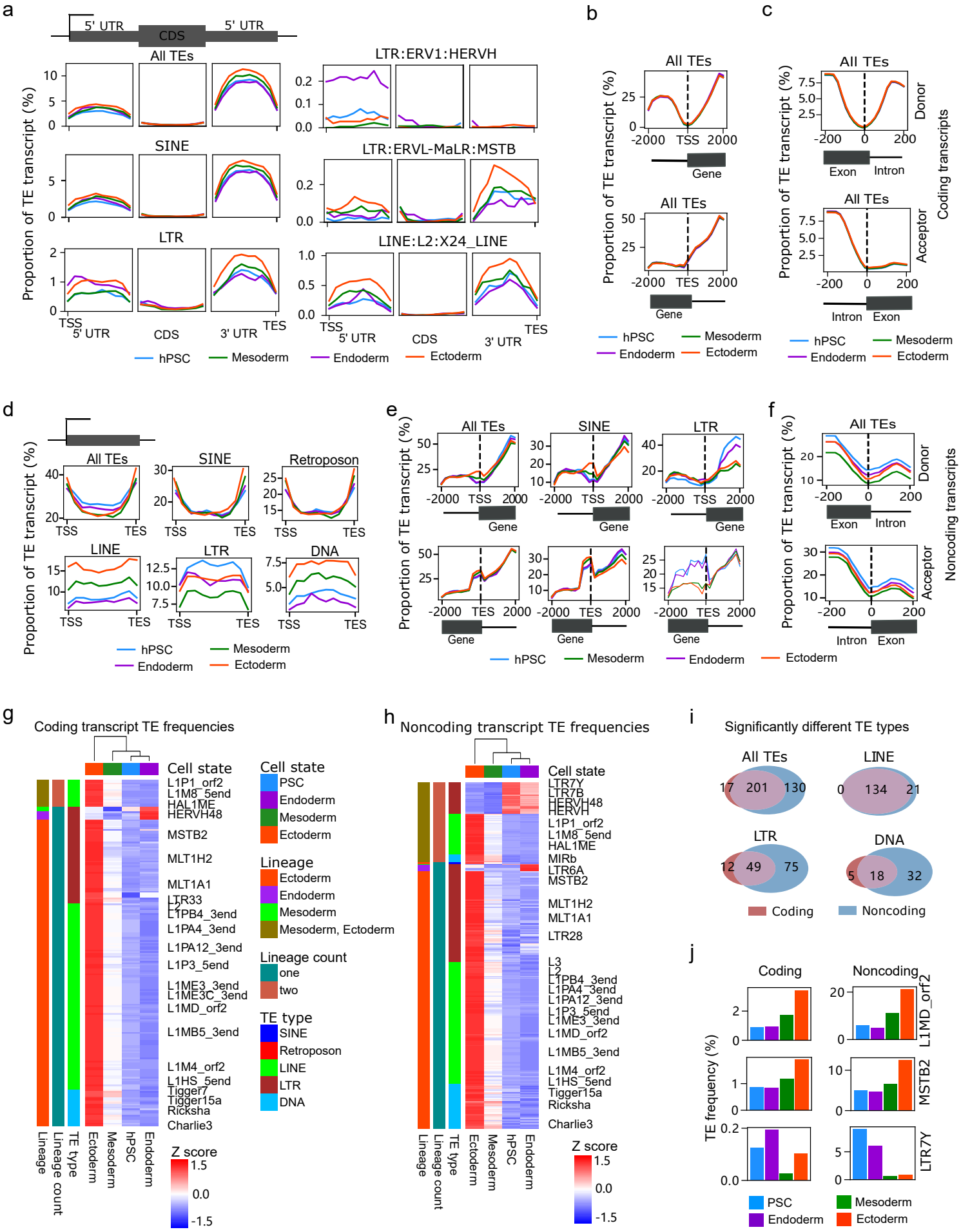
a
5' UTR
CDS
5' UTR
All TEs
LTR:ERV1:HERVH
SINE
LTR:ERVL-MaLR:MSTB
LTR
LINE:L2:X24_LINE
Proportion of TE transcript (%)
TSS
5' UTR
CDS
3' UTR
TES
hPSC
Mesoderm
Endoderm
Ectoderm
b
All TEs
Proportion of TE transcript (%)
-2000
TSS
2000
Gene
All TEs
TES
c
All TEs
Donor
Exon
Intron
Coding transcripts
Acceptor
d
All TEs
SINE
Retroposon
LINE
LTR
DNA
e
All TEs
SINE
LTR
f
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
Cell state
PSC
Endoderm
Mesoderm
Ectoderm
Lineage
Ectoderm
Endoderm
Mesoderm
Mesoderm, Ectoderm
Lineage count
one
two
TE type
SINE
Retroposon
LINE
LTR
DNA
Lineage
Lineage count
TE type
Ectoderm
Mesoderm
hPSC
Endoderm
Z score
1.5
0.0
-1.5
h
Noncoding transcript TE frequencies
Cell state
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
Mesoderm
Endoderm
Ectoderm

Fig. 4

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

991

992

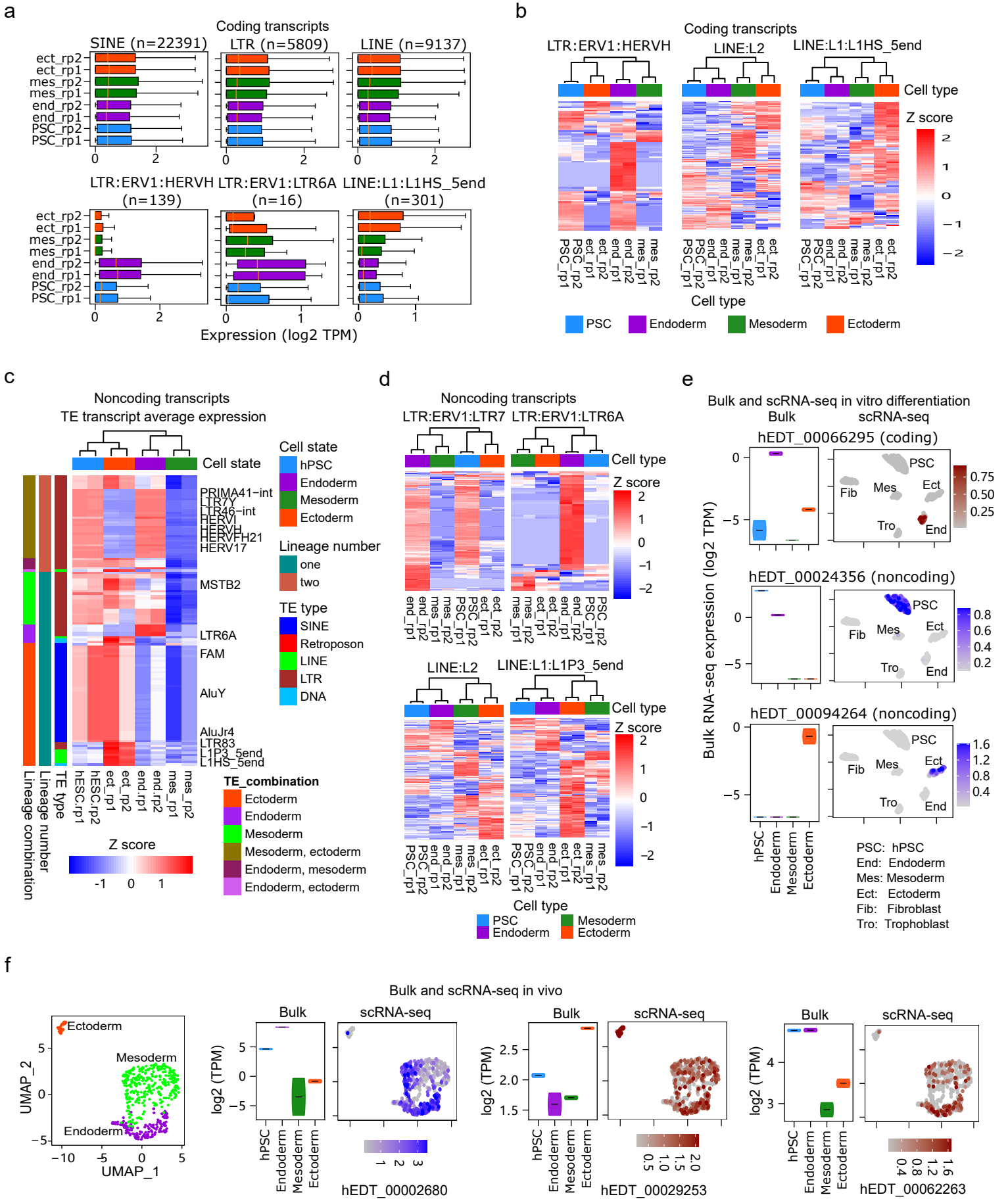
a
Coding transcripts
SINE (n=22391)
LTR (n=5809)
LINE (n=9137)
LTR:ERV1:HERVH (n=139)
LTR:ERV1:LTR6A (n=16)
LINE:L1:L1HS_5end (n=301)
ect_rp2
ect_rp1
mes_rp2
mes_rp1
end_rp2
end_rp1
PSC_rp2
PSC_rp1
Expression (log2 TPM)
b
Coding transcripts
LTR:ERV1:HERVH
LINE:L2
LINE:L1:L1HS_5end
Cell type
Z score
PSC
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

993 **Fig. 5: Dynamic expression patterns of TE-containing coding and noncoding transcripts**
994 **across different cell states. a**, Boxplots of the aggregate expression levels of coding transcripts
995 containing specific TEs and TE types. **b**, Heatmap showing the expression changes of
996 differentiation-induced differentially expressed coding transcripts containing the stated TEs. Each
997 row of the heatmap represents a transcript containing the indicated type of TE (top label). **c**,
998 Heatmap of aggregate changes in expression of noncoding transcripts containing different TEs.
999 Each row represents the mean of all transcripts containing a TE-type that is significantly different
1000 (adjusted p-value < 0.05 and fold change >=2) between hPSC and any of the differentiated cell
1001 states. For each TE, the lineage(s) with significant difference, and the number of lineage(s) with
1002 significant differences are shown. **d**, Heatmap showing the change in expression levels of
1003 differentiation-induced differentially expressed noncoding transcripts containing the indicated TEs.
1004 Each row represents a TE-containing transcript that is differentially expressed between hPSC and
1005 any of the three differentiated states. Bulk RNA-seq (left) and scRNA-seq (right) expression
1006 patterns of selected TE-containing transcripts in *in vitro* (**e**) and *in vivo* (**f**) differentiation data.

1007

1008

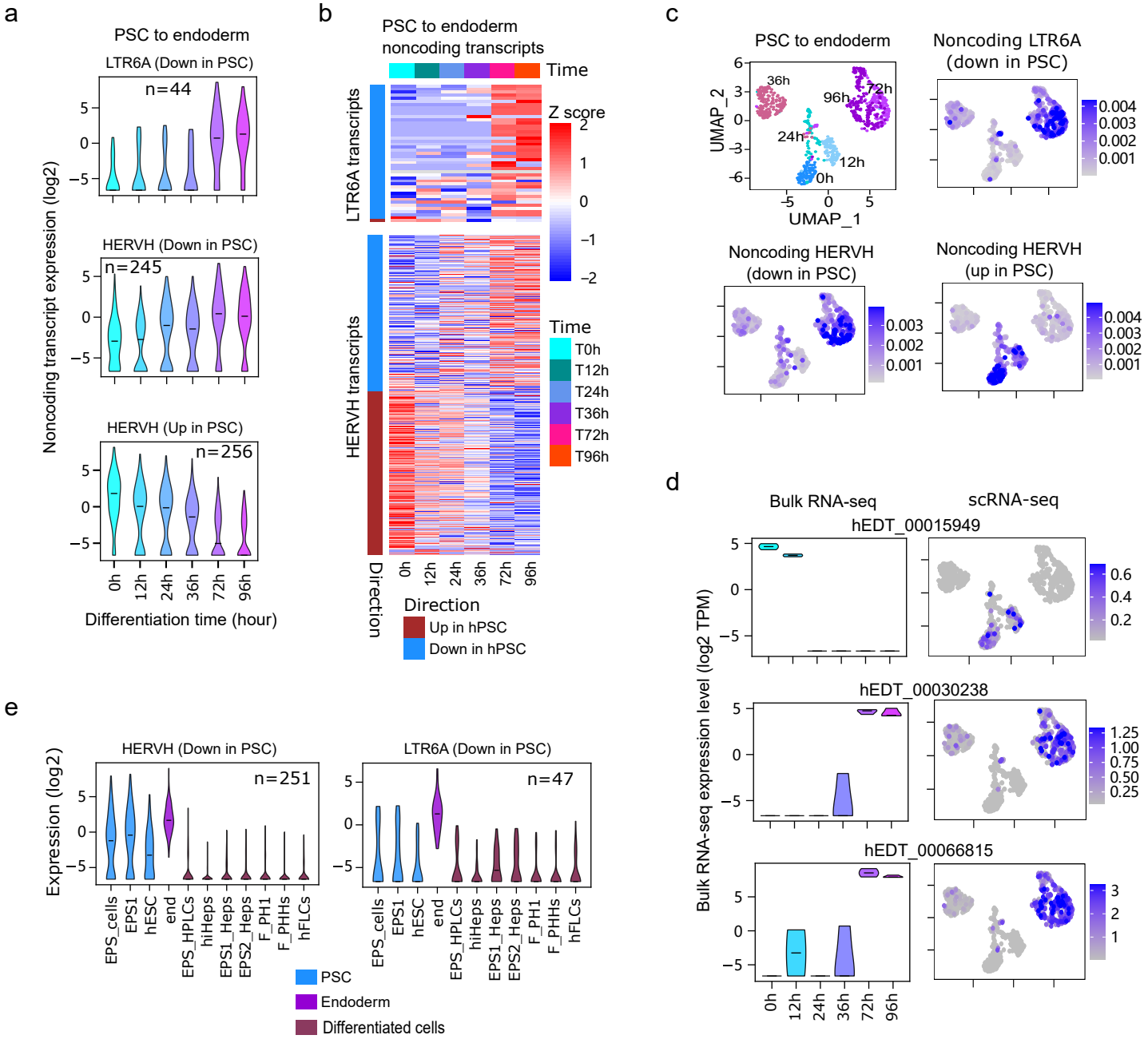
a
PSC to endoderm
LTR6A (Down in PSC)
n=44
HERVH (Down in PSC)
n=245
HERVH (Up in PSC)
n=256
Noncoding transcript expression (log2)
0h
12h
24h
36h
72h
96h
Differentiation time (hour)
b
PSC to endoderm
noncoding transcripts
Time
Z score
LTR6A transcripts
HERVH transcripts
Direction
Up in hPSC
Down in hPSC
T0h
T12h
T24h
T36h
T72h
T96h
c
PSC to endoderm
UMAP_2
UMAP_1
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

1009 **Fig. 6: Expression changes of TE-containing transcripts in endoderm lineage. a**, Violin plots
1010 of bulk RNA-seq expression levels of HERVH and LTR6A-containing noncoding transcripts
1011 during differentiation to endoderm. The data were from GSE75748. **b**, Heatmap showing the
1012 expression dynamics of HERVH and LTR6A noncoding transcripts differentially expressed in
1013 endoderm lineage. **c**, UMAP showing the single-cell expression data for a hPSC to endoderm time
1014 course. Hours of differentiation are shown. The three other UMAPs show example aggregate
1015 scores for transcripts containing the indicated TE types that are down-regulated in hPSCs or up-
1016 regulated in hPSCs. **d**, Expression dynamics in bulk RNA-seq (left) and scRNA-seq (right) of
1017 selected transcripts. **e**, Expression dynamics of endoderm-upregulated HERVH and LTR6A-
1018 containing noncoding transcripts in hepatocyte-related cells, using transcripts containing the
1019 indicated TE types, that are defined as down-regulated in hPSCs. (EPS: Extended pluripotent stem
1020 cells; EPS1: Stage 1 EPS; EPS2: Stage 2 EPS; hESC: Human ESC (H1); HPLC: Hepatic
1021 progenitor -like cells; EPS_HPLCs: EPS-derived HPLCs; hiHEPs: iPSC-derived hepatocytes;
1022 EPS1_Heps: EPS1-derived hepatocytes; EPS2_Heps: EPS2-derived hepatocytes; F_PH1: Fetal
1023 primary hepatocytes; F_PHHs: Fresh primary human hepatocytes; hFLC: Human fetal liver cells).

1024

1025

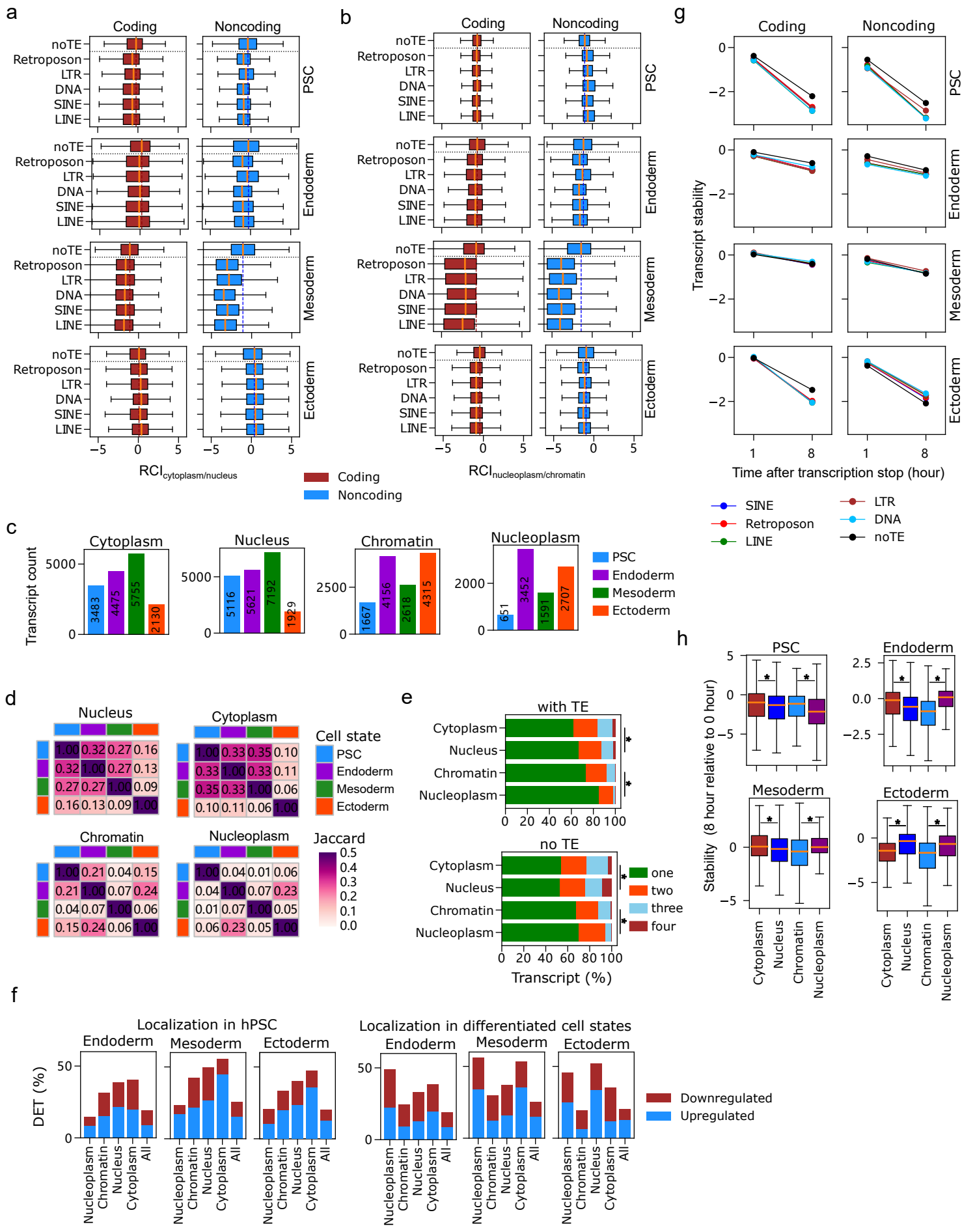
a
Coding
Noncoding
noTE
Retroposon
LTR
DNA
SINE
LINE
PSC
Endoderm
Mesoderm
Ectoderm
RCI cytoplasm/nucleus
Coding
Noncoding
b
RCI nucleoplasm/chromatin
g
Transcript stability
Time after transcription stop (hour)
SINE
Retroposon
LINE
LTR
DNA
noTE
c
Cytoplasm
Nucleus
Chromatin
Nucleoplasm
Transcript count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

Fig. 7

1026 **Fig. 7: Subcellular localization and stability of TE-containing transcripts**. **a**,
1027 Cytoplasm/nucleus relative concentration index (RCI). RCI is the log2-transformed relative
1028 expression in two tissues. **b**, Nucleoplasm/chromatin RCI. **c**, Bar charts showing the number of
1029 transcripts localized to different subcellular structures in different cell states using individual cell
1030 state assemblies. **d**, Heatmap showing subcellular localization consistency of hEDT transcripts
1031 localized to different structure across cell states. **e**, Stacked bars showing the distribution of the
1032 number of cell states in which transcripts are localized to a subcellular structure. (* represents Chi-
1033 square significance at p value < 0.05). **f**, The proportion of localized transcripts that were
1034 differentially expressed during differentiation process. The left charts show the localization in
1035 hPSC while the right charts show the localization in differentiated cell states. The distributions for
1036 all localizations were significantly different from the distribution of all hEDT set (Chi-square p
1037 value < 0.05). **g**, Line plots showing the transcript stability of TE-containing transcripts, using time
1038 from actinomycin D treated cells. The expression levels at time 1 and 8 hours, relative to the
1039 expression at 0 hour are shown for each TE type across different cell states. **h**, Boxplots showing
1040 the stability of transcripts that are localized to different subcellular structures across cell states (*
1041 represents Mann-Whitney test significance at p value < 0.05).