nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21808-x OPEN

Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE

Jiangping He[1], Isaac A. Babarinde[2], Li Sun[2], Shuyang Xu[3], Ruhai Chen[3,4], Junjie Shi[3], Yuanjie Wei[1], Yuhao Li[2], Gang Ma[2], Qiang Zhuang[2], Andrew P. Hutchins[2✉] & Jiekai Chen[1,3,4✉]

Transposable elements (TEs) make up a majority of a typical eukaryote's genome, and contribute to cell heterogeneity in unclear ways. Single-cell sequencing technologies are powerful tools to explore cells, however analysis is typically gene-centric and TE expression has not been addressed. Here, we develop a single-cell TE processing pipeline, scTE, and report the expression of TEs in single cells in a range of biological contexts. Specific TE types are expressed in subpopulations of embryonic stem cells and are dynamically regulated during pluripotency reprogramming, differentiation, and embryogenesis. Unexpectedly, TEs are expressed in somatic cells, including human disease-specific TEs that are undetectable in bulk analyses. Finally, we apply scTE to single-cell ATAC-seq data, and demonstrate that scTE can discriminate cell type using chromatin accessibly of TEs alone. Overall, our results classify the dynamic patterns of TEs in single cells and their contributions to cell heterogeneity.

[1] Center for Cell Lineage and Atlas (CCLA), Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, China. [2] Department of Biology, Southern University of Science and Technology, Shenzhen, China. [3] Key Laboratory of Regenerative Biology of the Chinese Academy of Sciences and Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China. [4] Joint School of Life Sciences, Guangzhou Medical University and Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China. ✉email: andrewh@sustech.edu.cn; chen_jiekai@gibh.ac.cn

Transposable elements (TEs) are a heterogeneous collection of genomic elements that have at various stages invaded and replicated extensively in eukaryotic genomes. The vast majority of TEs are fossils, and can no longer duplicate themselves, but they remain inside the genome and in mammals occupy nearly half the total DNA[1]. Intriguingly, it is becoming clear that both the active and remnant TEs are participating in evolutionary innovation and in biological processes[2–6], such as embryonic development[7–10], and in human disease and cancer[11,12]. Additionally, TEs carry cis-regulatory sequences and their duplication and insertion can reshape gene regulatory networks by redistributing transcription factor (TF) binding sites and evolving new enhancer activities[13–15]. TEs transcription also has a key influence upon the transcriptional output of the mammalian genome[16]. However, the role of TEs in cell type heterogeneity and biological processes has only recently begun to be explored in depth.

Single cell RNA-seq (scRNA-seq) has developed as a powerful tool to observe cell activity[17–19]. Many new techniques have been developed to recover or reconstruct missing observations, such as spatial, temporal, and cell lineage information. However, an important source of genomic information has so far been overlooked in single cell studies: the effect of TEs. Despite their importance, we lack quantitative understanding of how those genomic elements are involved in cell fate regulation at the single cell level. As TEs pose unique challenges in quantification, due to their degeneracy and multiple genomic copies, a prerequisite to understand TEs at the single cell level is a tool to quantify the hundreds to millions of copies of repetitive elements within the genome. To this end, we developed scTE, an algorithm that quantifies TE expression in single-cell sequence data.

In this study, we first demonstrate scTE's capabilities through an analysis of mouse embryonic stem cells (mESCs), which is one of the best characterized models for TE expression, as the expression of the endogenous retrovirus (ERV) MERVL marks a small population of cells in embryonic stem cell (ESC) cultures that are totipotent[20,21]. scTE accurately recovers the expected pattern of heterogeneous MERVL expression. Then, we apply our approach to several biological systems, including human in vitro cardiac differentiation, mouse gastrulation, adult mouse somatic cells, the induced pluripotent reprogramming process and human disease data. Overall, we gain insight into complex TE expression patterns in mammalian development and human diseases.

Results

Quantification of TE expression in single cells with scTE. Analysis of TEs pose special challenges as they are present in many hundreds to millions of copies within the genome. A common strategy in regular analyses is to discard multiple mapped reads, however, this leads to loss of information from TEs[22]. Assigning these reads to the best alignment location is the simplest way to resolve TE-derived reads, but it is not always correct for individual copies[22,23]. To solve this problem, we designed an algorithm in which TE reads are allocated to TE metagenes based on the TE type-specific sequence. Reads mapping to any TE copy in the genome are collapsed to a single TE subtype that represents that class of TE. The advantage is that errors in multimapping read allocation are minimized, the disadvantage is that TE genome location is lost. We built a framework named scTE with this strategy, scTE maps reads to genes/TEs, performs barcode demultiplexing, quality filtering, and generates a matrix of read counts for each cell and gene/TE (Fig. 1a and Supplementary Fig. 1a). scTE is easy to use, and its output is designed to be easily integrated into downstream analysis pipelines including, but not limited to, Seurat and SCANPY[24,25]. The algorithm can in principle be applied to infer TE activities from any type of single-cell sequencing-based data, like single-cell ATAC-seq data, DNA methylation, and other single-cell epigenetic data.

To evaluate the accuracy of scTE for non-TE gene expression, we compared gene expression from the standard Cell Ranger[26] pipeline, and the STARsolo[27] pipeline. scTE resulted in only minor changes in gene expression counts and high correlation (Pearson > 0.95) that is similar to the magnitude to the differences between STARsolo and Cell Ranger (Supplementary Fig. 1b). We then tested scTE's ability by in silico mixing two cells lines, MEFs (mouse embryonic fibroblasts) and ESCs in different ratios[28]. Comparison with the gene-based Cell Ranger pipeline[26], scTE shows nearly identical topology in a UMAP (Uniform Manifold Approximation and Projection) plot, and in marker genes expression (Fig. 1b and Supplementary Fig. 1c). Even when one cell type constitutes only a 1% minority in the mixture, scTE identified it correctly (Fig. 1b), indicating that scTE did not influence the global analysis of gene expression. These results demonstrate the sensitivity of scTE.

Next, we sought to explore TE expression, around 12–14% of the reads were derived from TEs (Fig. 1c). Requiring at least 2-fold change and FDR < 0.05, scTE detected 108 significantly differentially expressed TEs between ESCs and MEFs (Supplementary Fig. 1d), including ERVB7_1-LTR_MM, which is highly expressed in ESCs, and RMER10B in MEFs (Fig. 1d and Supplementary Fig. 1e). Furthermore, UMAP based on single cell TE expression alone could distinguish the cell types with the expected ratio (Fig. 1e), demonstrating TE expression discerns cell identity.

Deciphering TE heterogeneity in mouse ESCs and during human cardiac differentiation. It is known that a small subset of ESCs acquire a totipotent state named 2C-like cells and express a MERVL TE which also marks the embryonic 2-cell stage[20,29,30]. scTE could correctly identify this rare 2C-like subpopulation in UMAP plots, based on the specific marker genes *Zscan4c* and *Tcstv3*, and the expression of MERVL and MT2_Mm TEs (Fig. 2a, b and Supplementary Fig. 2a, b)[20,31]. If we discarded multiple mapped reads and only considered unique reads, the level of MERVLs was reduced, but it was still specifically expressed in the 2C-like cells (Supplementary Fig. 2c). Using TEs alone (no genes) the UMAP could correctly separate the rare 2C-like cells based on MERVL expression (Supplementary Fig. 2d). This confirms that scTE can correctly identify known TE patterns.

In humans, HERV-H LTRs are expressed in early embryos and human pluripotent stem cells (hPSCs), and contribute to pluripotency maintenance and somatic reprogramming[7,32–34], but little is known about TE expression dynamics during differentiation to somatic cells. Applying scTE to an scRNA-seq time series of hPSCs differentiating to cardiomyocytes[35], we observed the expected downregulation of HERV-H LTRs including LTR7 and HERVH-int during differentiation[4], concomitant with reduction in the expression of the pluripotency factor *POU5F1* (Fig. 2c, d and Supplementary Fig. 2e). During in vitro cardiac differentiation of hPSCs there is a bifurcation towards definitive cardiomyocytes (dCM) and non-contractile cells (Fig. 2c). Between these two branches, marked by *NKX2-5* and *SPARC*, respectively, we found differential expression of TEs such as LTR32, MER57A-int and MER45A in the dCM cells, whilst, MLT1H1, HERVIP10B-int and LTR5A were specifically expressed in the non-contractile cells (Fig. 2e, f and Supplementary Fig. 2f). Independent bulk RNA-seq data[36] demonstrated that these TEs were expressed in late cardiac differentiation

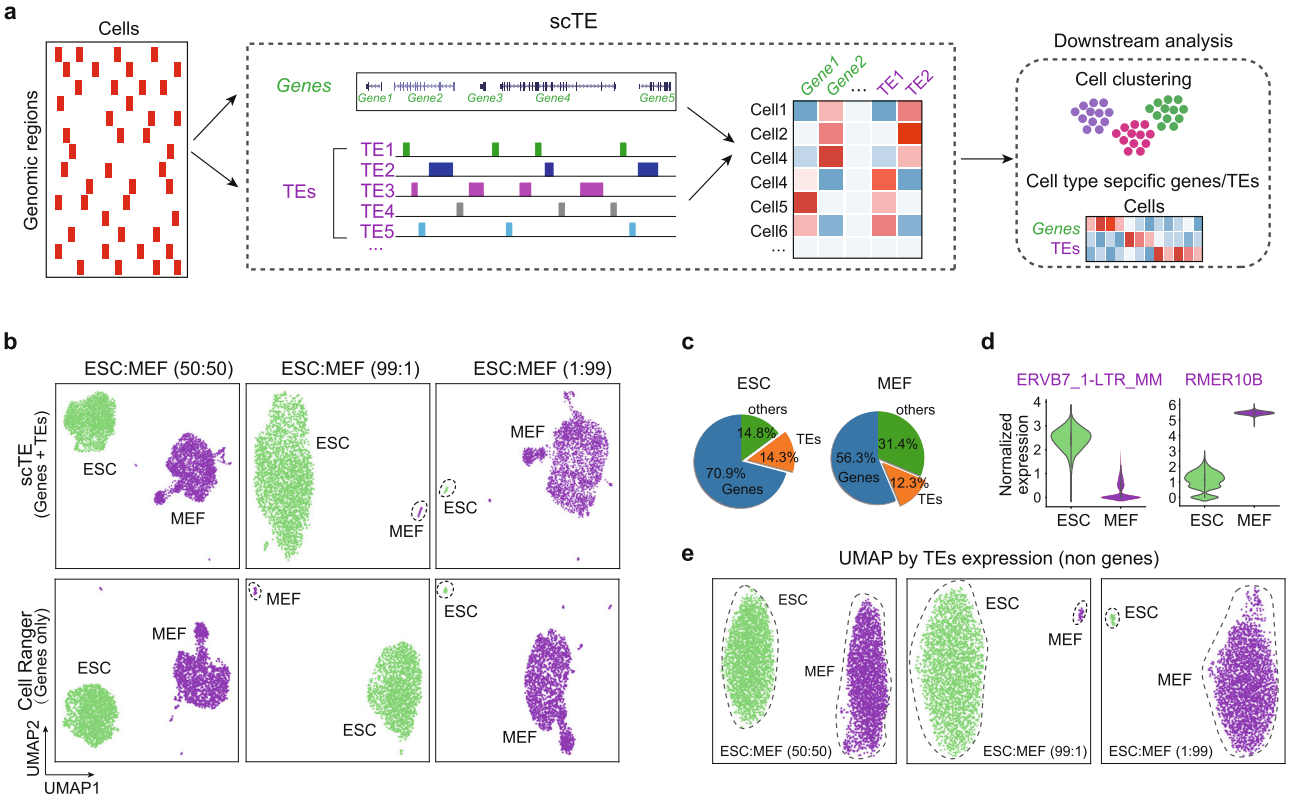

Fig. 1 scTE workflow and applications. a Schematic of the workings of scTE. For scRNA-seq data the reads are mapped to the genome, and assigned to either a gene, or a metagene model of a TE. Multimapping read data will assign the best mapping read to a type of TE. Reads are always mapped to a gene first, and then a TE if no gene is found. The resulting assignments are then collapsed into a matrix of read counts for each cell, versus each gene/TE. This matrix can be used in downstream applications. Genes are colored in green and TEs are colored in purple in all figures. **b** UMAP plot showing mixtures of MEFs and ESCs in the indicated ratios. The top panels show scTE analysis, the lower panels show Cell Ranger analysis results. Cells are colored by their sample of origin. **c** Percentage of reads mapping to genes, TEs or other regions of the genome in MEFs and ESCs. **d** Violin plot showing the expression of selected TEs in MEFs and ESCs. **e** As in (**b**), but only TE expression was used.

(Supplementary Fig. 2g), however, as the bulk is a mixture of dCM and non-contractile cells, the restriction of these TEs to divergent fates can only be observed in the scRNA-seq data. This highlights the importance of analyzing TE expression in sc-RNA-seq data, as MLT1H1 is very high in the bulk RNA-seq, but this hides the reality that it is restricted to the non-contractile cells and plays no role in dCMs (Fig. 2e, f and Supplementary Fig. 2g). To explore if the reads are derived from relatively intact ERV elements or truncated fragments, we compared the expression correlation between LTR and their internal ERV sequence. LTR7 and HERVH-int were strongly correlated (Pearson ~0.8; Supplementary Fig. 2h and Fig. 2d), and LTR5A and LTR6A were also positively correlated to their internal ERVs (Supplementary Fig. 2h,i), indicating their expression may be from relatively intact elements. We also noticed that some LTR expression did not correlate with their ERV, such as LTR32, which is specifically expressed in the dCMs, while its internal HERVL32-int is not expressed in any cell types (Supplementary Fig. 2h,i), this suggests a disconnect between the expression of the LTR and ERV, and hints at separate regulation or truncation of the LTR/ERV pair.

Analysis of TEs in mouse gastrulation and early organogenesis identifies the widespread cell fate-specific expression of TEs. The previous analysis showed TE expression dynamics during in vitro cardiac differentiation, next we explored complex in vivo developmental processes. TE expression is dynamic during pre-implantation development[7], however, the expression of TEs in gastrulation has not been described. We took advantage of the single-cell time course of mouse gastrulation[17]. Analysis with scTE did not introduce any unexpected sample-bias, and a side-by-side comparison could retrieve similar patterns of marker gene expression in the expected lineages (Fig. 3a and Supplementary Fig. 3a–f). We found every lineage expressed a series of lineage-specific TEs (Fig. 3a, b, and Supplementary Fig. 4a–c). In the extraembryonic ectoderm cells, IAP and RLTR45-family TEs were activated (Fig. 3b, c), and in *Apoa2*+ extraembryonic endoderm cells, MER46C, RLTR20B3, and LTRIS2 were upregulated (Fig. 3b, d). The expression of these TEs was validated using bulk RNA-seq from in vitro[37–39] mimics of these embryonic stages, including ESCs, epiblast stem cells (EpiSCs), extraembryonic endoderm cells (XENs) and trophoblast stem cells (TSCs) (Fig. 3e). Other embryonic lineages, particularly the *Gypa*+ erythroid and the *Tnnt2*+ cardiomyocyte lineages expressed specific TEs such as L1_Mur and L1ME3D, respectively (Fig. 3b, f).

As this dataset provides dynamic trajectories for each lineage, we wondered if TEs were transiently activated during cell fate transitions. To this end, we noticed ETnERV3-int, whose expression coincides with the early development of the cardiac fate from the mesoderm, and is reduced in *Tnnt2*+ cells, while L1ME3D was expressed in the *Tnnt2*+ cells (Fig. 3g). Consistently, ETnERV3-int was specifically expressed in in vitro derived cardiomyocytes, which more closely resemble a fetal state, whilst L1ME3D was expressed only in the mature heart (Fig. 3h)[40,41]. However, the bulk samples could not capture the complexity of the transient expression of ETnERV3-int which extended from the late epiblast into the endoderm and mesoderm. To expand on this,

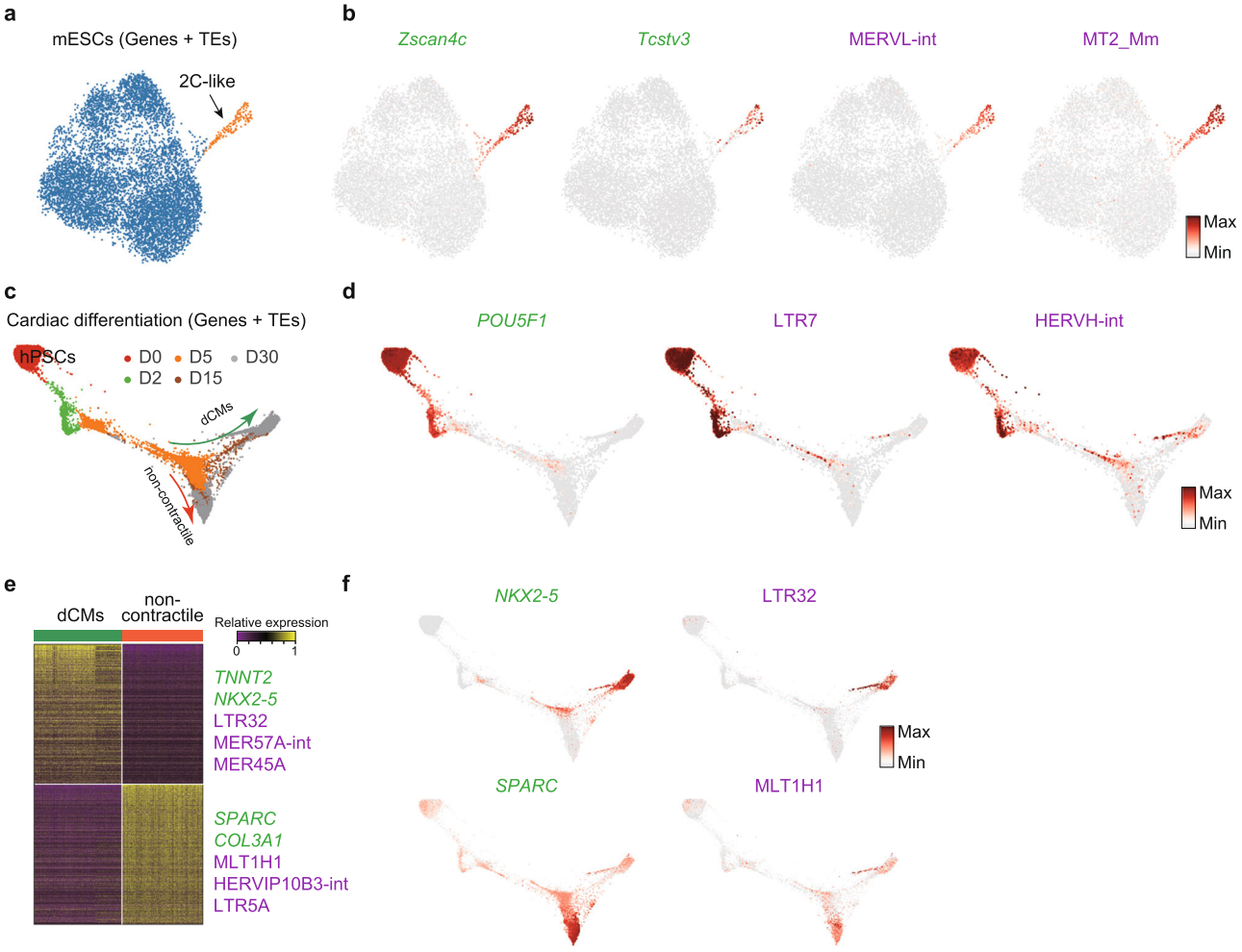

Fig. 2 Dynamic transcription of TEs in ESCs and during cardiac differentiation. a UMAP plot of mouse ESCs. Cells are colored by cell type cluster. **b** Same as (**a**), but cells are colored based on the expression of the indicated genes and TEs. *Zscan4c* and *Tcstv3* are marker genes for the 2C-like cells. **c** Trajectory reconstruction of single cells through a cardiac differentiation timecourse showing the definitive cardiomyocytes (dCMs) branch and noncontractile branch. Days of differentiation (D) are labeled. **d** As in (**c**), but cells are colored by the expression of the indicated genes and TEs. **e** Heatmap of expression differences between dCM (contractile) branch and noncontractile branch cells, selected differentially expressed genes and TEs are labeled. **f** As in (**d**), but cells are colored by the expression level of the indicated genes and TEs.

we reanalyzed an scRNA-seq dataset of the developing mouse embryonic heart[42] (Fig. 3i and Supplementary Fig. 5a–c), and found that ETnERV3-int was expressed in the myocardium and epicardium, but not in the endocardium, neural crest, and embryonic cells (Fig. 3j). L1ME3D was expressed in *Tnnt2*+ myocardium, however, in an inverse pattern with respect to ETnERV3-int (Fig. 3j, k). Therefore, ETnERV3-int is expressed in an intermediate stage of cardiac lineage development. Intriguingly, there was a close relationship between the expression of ETnERV3-int and *Isl1* gene, which marks multipotent progenitors[42] (Fig. 3j). These results highlight the complex patterns of TE expression in developmental processes.

Widespread tissue-specific expression of TEs in somatic cells. As we detected heterogeneity of TE expression during organogenesis and cardiac differentiation, we next took advantage of scRNA-seq to explore TE expression heterogeneity in somatic tissues. As we revealed unexpected heterogeneity of TEs in somatic MEFs and during organogenesis, we next measured TE expression in somatic cells using the *Tabula Muris* large scale scRNA-seq dataset that profiles 20 mouse organs[43] (Fig. 4a). Surprisingly, our analysis identified in total 130 TEs that were specifically expressed in distinct cell types (Fig. 4b and Supplementary Fig. 6a). These associations include the expected expression of LINE1 elements in brain cells, of which many L1 family members like L1MEh, L1M, L1MC4a, L1MA7, and L1P5 elements are specifically expressed in oligodendrocytes or microglia (Fig. 4c and Supplementary Fig. 6a). We also found expression of LTR58, MLT1EA-int, MER110, and RLTR46 specifically in B cells, T cells, type B pancreatic cells, and hepatocytes, respectively (Fig. 4c).

TE expression is regulated by chromatin modification and transcription factors (TFs)[3], thus, we wondered if we could infer the regulatory network between TFs and TEs from large scale scRNA-seq data, taking advantage of the improved cell type definitions from the scRNA-seq data. The co-expression relationships often reflect biological processes in which many genes with related functions are coordinately regulated. Therefore, we reasoned that if a TE is regulated by a TF, they should be co-expressed. To identify TF–TE regulatory relationships, we performed co-expression analysis, and identified the specific co-clustering of neural genes and TEs (*Sox2* and *Olig1*), the immune system (*Cebpe, Tcf7, Pax5,* and *Sall1*), the endoderm/pancreas (*Gfi1b, Nkx6-1*, and *E2f8*), and other lineages (Fig. 4d, e and Supplementary Fig. 6b). Motif analysis showed that the SOX2

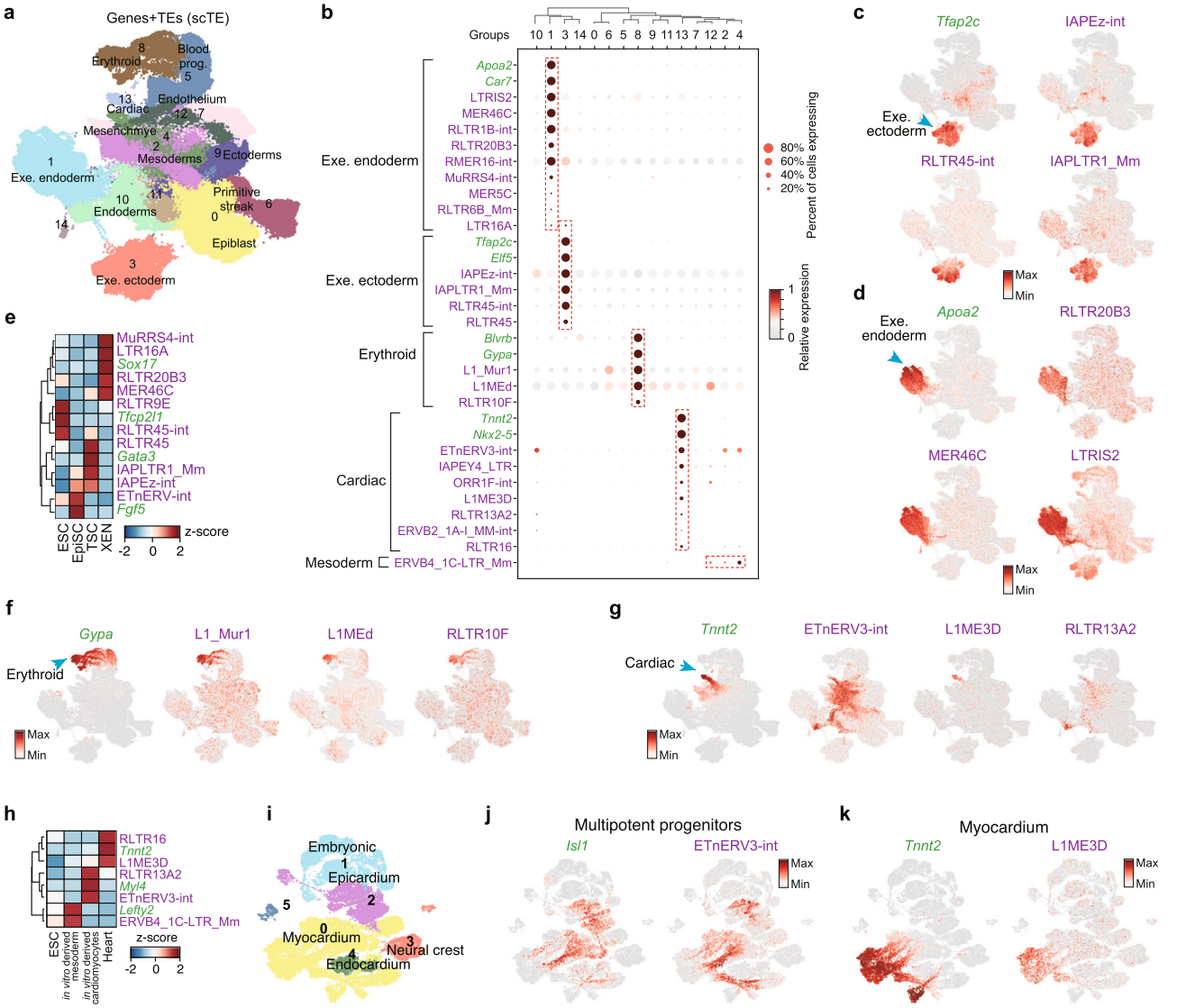

Fig. 3 Widespread cell type-specific expression of TEs during gastrulation. a UMAP plots of the mouse gastrulation data using both genes and TEs. Selected lineages are labeled (Leiden, resolution = 0.3). **b** Dot plot showing a selection of marker genes and TEs for the indicated cell lineages. **c** Expression of the indicated extra embryonic ectoderm gene *Tfap2c* and selected TEs. **d** Expression of the extra embryonic endoderm marker gene *Apoa2* and selected TEs. **e** Expression of the indicated TEs and marker genes in bulk RNA-seq data from ESCs, EpiSCs, XEN (extra embryonic endoderm cells) and TSCs (trophoblast stem cells). *Tfcp2l1*, *Fgf5*, *Gata3*, and *Sox17* serve as markers for ESCs, EpiSCs, TSCs, and XEN cells, respectively. Data are displayed as a *z*-score using the variance from all genes. **f** Expression of the erythroid marker gene *Gypa*, and selected TEs. **g** Expression of the cardiac marker gene *Tnnt2* and selected TEs. **h** Expression of the indicated TEs and marker genes from bulk RNA-seq data. **i** UMAP plot of the embryonic mouse heart scRNA-seq data using both TEs and genes. The indicated developmental stages are labeled as in the original study. **j**, **k** UMAP as (**i**), but cells are colored by the expression of indicated genes/TEs.

motif was significantly enriched within RLTR13F TEs (Supplementary Fig. 6c). ChIP-seq data analysis also demonstrated the binding of the TFs TCF7[44], SOX2[45], and TFAP2C[46] to RLTR10D2, RLTR13F and RLTR13D5 TEs, respectively (Fig. 4f). These results highlight the deep link between TE and TF activity, indicating those TFs may be responsible for activating TEs in the corresponding cell types.

We next explored two cell lineages where TE activity is known to be involved, the neural and immune cell lineages[47–49]. TEs have contributed both exapted proteins, enhancer sequences, and non-coding RNAs to regulate innate immune responses[47,48]. In the neural system, LINE TEs are especially active and whilst their activity remains unclearly understood they are deregulated in many neurological disorders[49]. Subgrouping the cells from microglia and neuron samples identified several distinct cell types (Supplementary Fig. 7a–c), within which cell type-specific expression of TEs was observed (Supplementary Fig. 7d, e). Next, with the pooled immune cells from marrow, spleen, and thymus, 12 distinct immune cell subtypes were defined (Supplementary Fig. 7f, g). Intriguingly, besides finding additional cell type-specific TEs in T cells, B cells and granulocytes, a series of TEs were restricted to subtypes of T cells and B cells (Supplementary Fig. 7h and i). These data show different degrees of subtype specific signatures of TEs in the neural and immune system, and highlight the importance of looking beyond only genes when exploring how those systems differ.

TEs are activated during somatic cell reprogramming, in a heterogonous and cell branch restricted manner. The above analysis has revealed the well-ordered dynamic expression of TEs

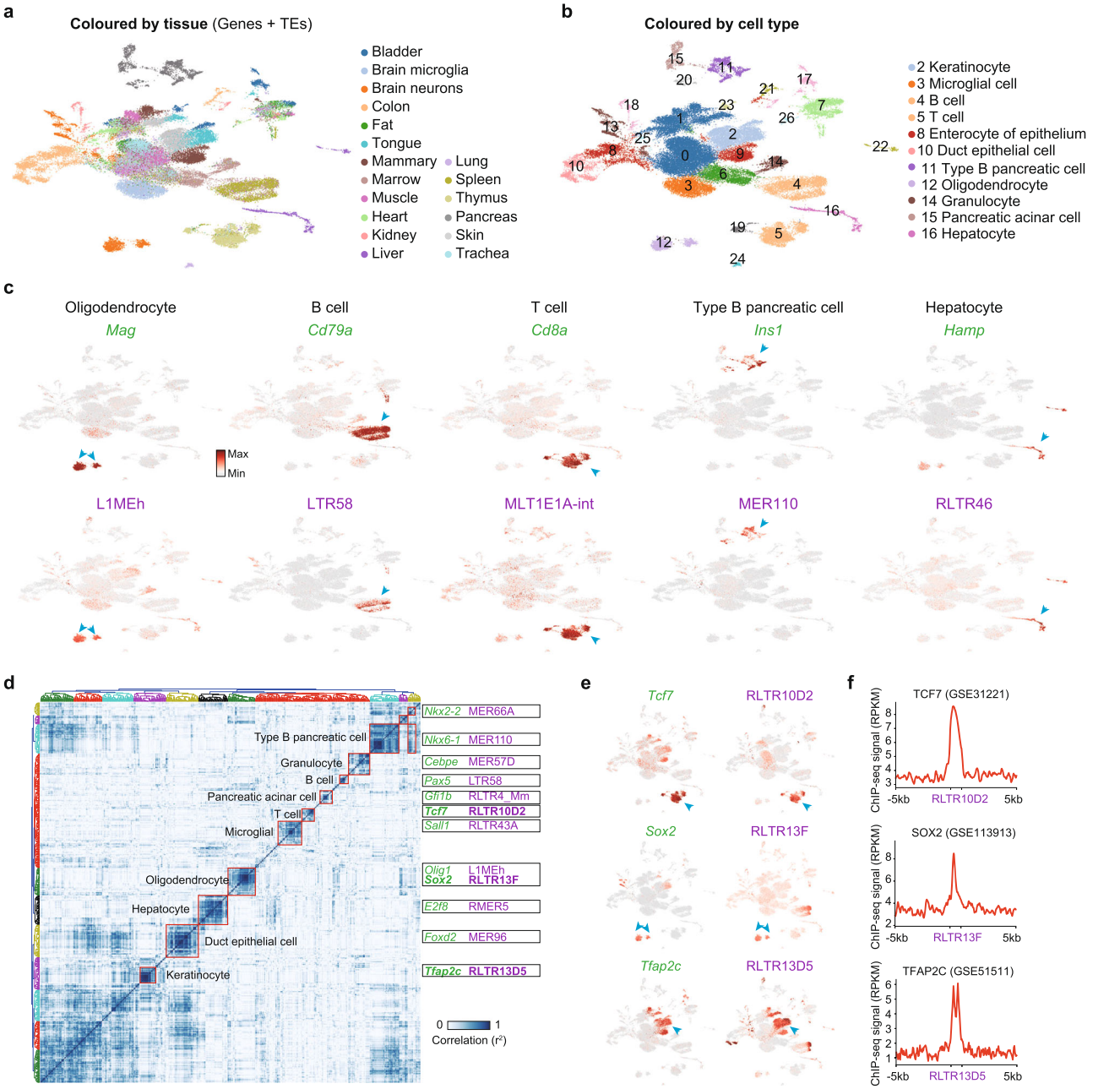

Fig. 4 Class-specific expression of TEs in somatic cells. a UMAP plots of the *Tabula Muris* data, using both genes and TEs as analyzed with scTE. The tissue sources for the cells are indicated. **b** UMAP plot as in (**a**), but clustered into groups (Leiden, resolution = 0.5). **c** Same as (**b**), but cells are colored by the expression of indicated genes/TEs. **d** Correlation heatmap showing the co-expression of TFs and TEs. **e** UMAP plots showing the expression of indicated TFs and TEs. **f** Read count tag density pileups for TCF7, SOX2, and TFAP2C ChIP-seq data on the indicated TEs.

in developmental processes, we then wondered if TEs undergo similar stage-specific regulation during somatic reprogramming. Somatic cells can be reprogrammed to induced pluripotent stem cells (iPSCs) by various methods, such as ectopic expression of a group of pluripotency transcription factors[28,50,51], or cocktails of chemicals[52,53]. The reprogramming process is highly heterogeneous, with abundant non-reprogramming cells and divergent cell fate transition routes[28,54]. We took advantage of reprogramming scRNA-seq data to investigate the expression of TEs during these drastic cell fate transitions. Reprogramming induced by *Oct4/Pou5f1*, *Klf4*, *Sox2*, and *c-Myc* (OKSM) generates detectable intermediate branches, including iPSCs, trophoblast, stromal and neural-like cells (Fig. 5a and Supplementary Fig. 8a–d)[54]. We identified specifically expressed TEs in each cell branch (Supplementary Fig. 8a–d). For example, the TEs ERVB7_1-LTR_MM, IAPEz-int, RLTR4_Mm, and Lx were specifically expressed in iPSCs, trophoblast, stromal and neural-like branches, respectively (Fig. 5b). ERVB7_1-LTR_MM (MusD) and IAPs are upregulated during reprogramming[55], however using scRNA-seq data we show that only ERVB7_1-LTR_MM, as well as ETnERV-int and RLTR13G, were upregulated in the successful reprogramming route, initiating at the mesenchymal-to-epithelial transition (MET) and peaking at the iPSCs stage (Fig. 5b and Supplementary Fig. 8a). In contrast, the trophoblast-branch expressed IAPEz-int and IAPLTR1_Mm (Fig. 5b and Supplementary Fig. 8c), which are also expressed in in vivo extra embryonic ectoderm cells (Fig. 3c), suggesting consistent regulation between development and reprogramming.

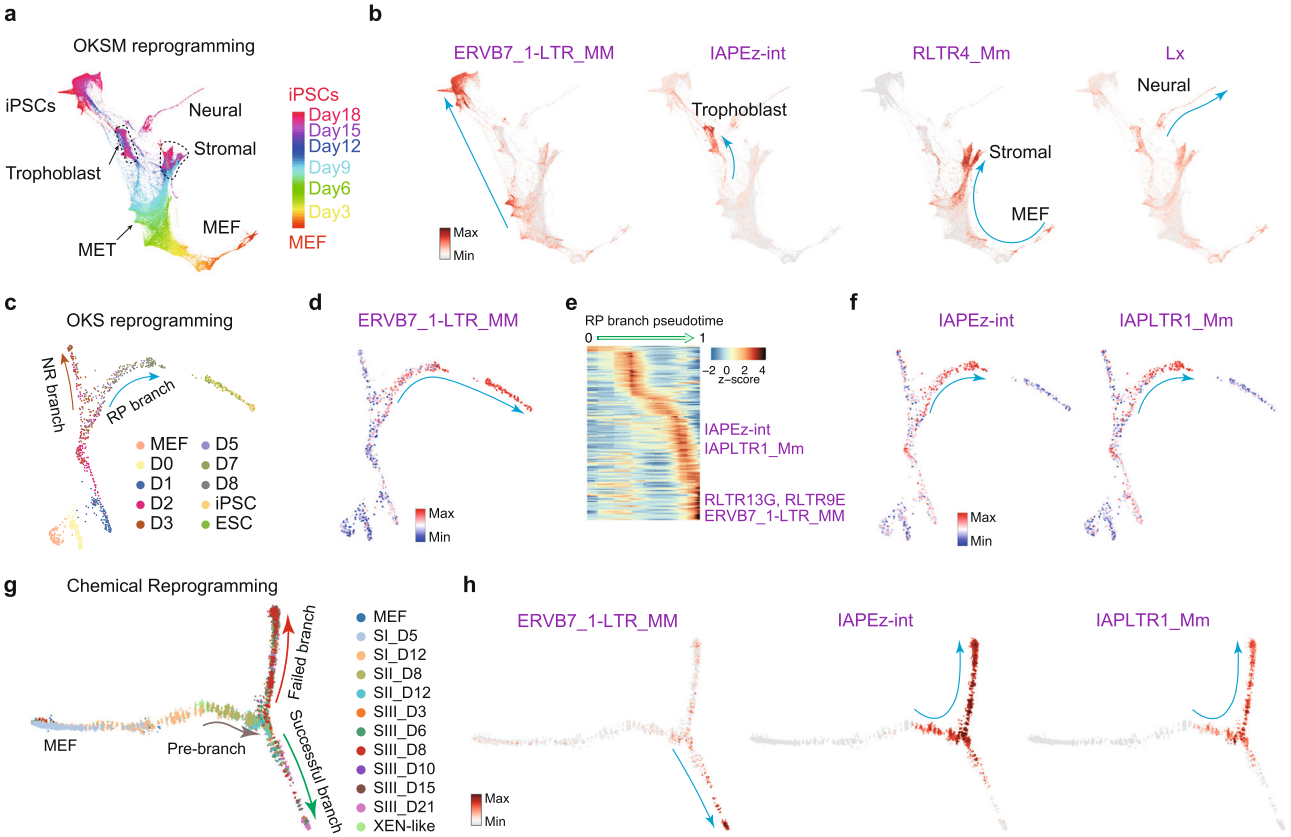

Fig. 5 Stage-specific expression of TEs in somatic cell reprogramming. a Trajectory reconstruction during OKSM reprogramming, cells are colored by time point. **b** As in (**a**), but cells are colored by the expression of the indicated TEs. **c** Force-directed (FR) layout of cells during OKS reprogramming, cells are colored by time point. **d** Same with (**c**), but cells are colored by the expression change of the ERVB7_1-LTR_MM TE during reprogramming. **e** Expression heatmap of the top 145 dynamically expressed TEs in a pseudotime ordering for the RP branch, selected TEs are indicated. **f** Expression changes of the indicated TEs during reprogramming. **g** Trajectory reconstruction during chemical reprogramming, cells are colored by time point. **h** As in (**g**), but showing TE expression specific to the successful or failed branches of reprogramming.

We then analyzed reprogramming induced by *Oct4*, *Klf4*, and *Sox2* (OKS)[28] or only chemicals[31]. There are two validated branches during OKS-mediating reprogramming[28] (Fig. 5c), and we found many TEs, such as ERVB7_1-LTR_MM, that were specifically upregulated in the reprogramming-potential (RP) branch, and were excluded from the non-reprogramming (NR) branch (Fig. 5d and Supplementary Fig. 9a). As OKS reprogramming data was generated with both the 10x (3' biased) and C1 (full-length) methods, we took advantage of these matching datasets to compare the influence of the single-cell RNA-seq protocol. Broadly, they matched well between each other for both genes and TEs (Supplementary Fig. 9b), and we could detect similar patterns of TE expression in both the 10x and C1 (Supplementary Fig. 9c, d). However, as the 10x results in considerably more cells than the C1 platform, a unique cell type "neuron-like" (NL) could only be detected in the 10x data, and these cells expressed LINE1 elements (Supplementary Fig. 9c, d). IAPEz-int and IAPLTR1_Mm were expressed in the RP branch but were ultimately silenced in the reprogrammed cells (Fig. 5e, f), suggesting IAPs were only activated in a pre-reprogrammed state and are down-regulated before the finalization of reprogramming. Bulk RNA-seq can identify overall changes in TE expression, however, the dynamics and branch-restricted TE expression can only be observed from the scRNA-seq. We validated the expression of ERVB7_1-LTR_MM and IAPs by qRT-PCR (Supplementary Fig. 9e), demonstrating that IAPs are silenced in ESCs. Similar to OKS-mediated reprogramming, chemical-mediated reprograming bifurcates into two branches (Fig. 5g and Supplementary Fig. 9f)[31], and TEs, marking an intermediate 2C-like program, were activated at the root of the successful branch (Supplementary Fig. 9g, h). ERVB7_1-LTR_MM and RLTR13G were specifically upregulated in the successful branch, whilst IAPEz-int and IAPLTR1_Mm were activated in the pre-branch and failed branch (Fig. 5h and Supplementary Fig. 9i, j).

The three reprogramming systems described above can progress along different paths to reprogramming[28,31,54], however, the same TEs are regulated in similar patterns in the three systems, suggesting common regulatory mechanisms for TEs. Indeed, we found IAPLTR1_Mm TEs are rich in DNA-binding motifs for JUN and IRF2 (Supplementary Fig. 9k), whose expression closely matched IAP expression in all three reprogramming systems (Supplementary Fig. 9l) and are known to impair reprogramming[56,57]. This suggests that their downregulation deactivates the IAPs before the finalization of reprogramming. Overall, these results unveiled a deeper unappreciated role of dynamic TE expression in iPSC formation.

Passive transcription is not a major contributor to TE expression in single cells. We next evaluated the effects of TE expression from passive co-transcription with genes, especially TEs that are retained in transcribed introns[58,59]. First, we observed that the 10x data are significantly 3′ biased, with most read counts in the 3′ end of genes, and a very low tag density for the gene body (Supplementary Fig. 10a), indicating read-through across the gene body is not a major part of the expression

measure. Nonetheless, to rule out a major influence of intronic TEs on determining cell type-specific TEs, we performed TE counts using only reads from outside gene bodies (using the nointronic mode in scTE). Analysis of the cell type-specific TEs between MEFs and ESCs in the default mode (exclusive) (Supplementary Fig. 1d), indicated that the majority of those cell type-specific TEs remained specific in the nointronic mode (94/108), and just 14 TEs were altered in the nointronic mode (Supplementary Fig. 10b). To explore the impact of genomic proximal genes to those cell type "inconsistent" TEs, we collected cell type-specific genes (Supplementary Fig. 10c), and then compared these genes with the locations of those cell type "inconsistent" TEs. We did not detect any significant genomic proximal correlation between them (Supplementary Fig. 10d), indicating that intronic read counts from high expressed genes is not a major issue for TE analysis in 10x data from whole-cell scRNA-seq. Potentially this may be more of an issue in nuclear scRNA-seq, where intron retention is more common. Nonetheless, we noticed some cell type-specific expressed TEs that are inside the intron of a gene that was not expressed (Supplementary Fig. 10e), indicating that the removal of intronic reads may bias TE quantification for some TEs in single cells. We next expanded this analysis to the *Tabula Muris* atlas dataset performed using the C1 platform. Similarly, most cell type-specific TEs did not correlate with cell type-specific genes, but there were a limited number of correlations (11/129), especially for LTR90A, RLTR19B etc. (Supplementary Fig. 10f–h). Above all, these results suggesting that a relationship between genes and proximal TEs occurs, but only in a minority of cases.

To evaluate the influence of TE mappability on quantification accuracy, we performed cross correlation analysis between read mappability, read counts and the coefficient of variance (CV) for each TE sub-type, and show there was no obvious correlation, and generally the cell type-specific expressed TEs with high variance (CV high) had a high mappability score (Supplementary Fig. 10i, j), indicating that the cell type specific TEs identified by scTE are reliable.

Inferring TE-associated accessibility from scATAC-seq data. Beyond scRNA-seq, many other single-cell sequencing techniques[60–62] have shown great potential to explore cell heterogeneity, and increased insight could be fueled by the additional information provided by scTE. For instance, we reasoned that scTE would be informative for the analysis of scATAC-seq data and potentially other single-cell epigenetic data because TEs have a wide array of chromatin states[3], are widely bound by transcription factors[63], and can act as enhancers[15] (Fig. 6a). We then applied scTE to a dataset of fluorescence-activated cell sorted (FACS) mouse cells[64], including cardiac progenitor cells (CPCs), $CD4^+$ T cells, ESCs and skin fibroblasts (SFs). Intriguingly, scTE could accurately recover the expected cell types, based on only the reads that mapped to TEs (Fig. 6b). Specific accessibility of RLTR13A, RLTR4_Mm, RLTR13G and RMER19B/C was found in the CPCs, CD4+ T cells, ESCs and SFs, respectively (Fig. 6c, d and Supplementary Fig. 11a). And motif enrichment of these cell-type specific TEs identified known master regulators of these cell types, such as GATA4/HAND1/T for CPCs, ETS1/TCF3 for T cells, SOX2/POU5F1/NR5A2 for ESCs and FOS/MAF for SFs (Supplementary Fig. 11b), indicating these TEs may act as cis-regulatory elements bound by transcription factors. For instance, scTE identified an RLTR13A TE within an intron of *Smyd1*, a gene essential for heart development[65–67], which was specifically open in CPCs (Fig. 6e), and was specifically expressed in the myocardium of the fetal heart (Fig. 6f). The above dataset was presorted, which meant there was a priori information about the cell type. In a more challenging case, we analyzed single-cells from unsorted mouse spleen[64]. In this scATAC-seq dataset with scTE we can detect the major spleen cell types, including B cells, macrophages (Mq), granulocytes, natural killer (NK) cells and T cells, based on accessibility at known cell type-specific genes (Supplementary Fig. 11c, d). Additionally, each cell type had specifically opened TEs (Supplementary Fig. 11e). Finally, we applied scTE to scATAC-seq data of human primary cells, of a peripheral blood monocyte (PBMC) population, and could recover the major cell types and cell type-specific TEs (Supplementary Fig. 11f–i), which could be validated by independent bulk ATAC-seq data from FACS sorted cells (Supplementary Fig. 11j)[68]. These results indicate that quantifying chromatin accessibility on TE regions is informative for characterizing cell types and may assist the problems posed by scATAC-seq analysis due to its especially sparse nature[69].

Disease-specific expression of TEs. The unexpected widespread TE heterogeneity amongst embryonic and somatic cell types raised the question as to whether there is TE heterogeneity in diseased cells. Alzheimer's disease (AD) is an age-associated neurodegenerative disorder that is characterized by progressive memory loss and cognitive dysfunction for which there is no known cure. TEs have been reported to be highly active during aging and may contribute to age-dependent loss of neuronal function[70]. To explore the expression of TEs in AD, we reanalyzed the scRNA-seq data from a mouse model of AD expressing five human familial AD gene mutations, which contained 13,114 single cells with age and sex-matched wild-type (WT) controls using the MARS-seq platform[71] (Fig. 7a). Projecting the cells with a UMAP, we recovered the major groups of cells in AD and WT, including the unique disease-associated microglia cluster cells (M2) identified in the original study (Fig. 7b and Supplementary Fig. 12a). Differential expression analysis demonstrated significant changes in gene expression in M2, including previously described AD risk factors such as *Apoe*, *Tyrobp*, *Lpl*, *Cstd*, and *Trem2* (Fig. 7c and Supplementary Fig. 12b). Intriguingly, we also found many TEs such as ERVB7_2-LTR_MM, RLTR17, RLTR28 and Lx4B that were significantly higher and specifically expressed in M2 (Fig. 7c, d and Supplementary Fig. 12c), indicating those TEs may also be involved in AD development.

Type 2 diabetes (T2D) is a common human disease caused by a combination of increased insulin resistance and reduced mass or dysfunction of pancreatic beta cells. We reanalyzed scRNA-seq from two independent studies of the human pancreas in healthy and T2D individuals[72,73]. The major cell types in the pancreas, including alpha, beta, gamma/PP, and delta cells clustered without a visible disease-specific pattern, indicating no drastic change in cell type (Fig. 7e and Supplementary Fig. 12d). Contrasting the transcriptome from healthy and T2D in each cell type independently, *CD36* and *DLK1* was upregulated in T2D alpha and beta cells respectively (Fig. 7f), as reported by the original studies[72,73]. Notably, many TEs were significantly highly expressed in T2D beta cells, including L1MC, L1MA4A, Tigger3a, MLT2B4. This differential expression pattern was near identical between the two independent datasets (Fig. 7f). Critically, none of these observations could be observed using bulk RNA-seq datasets (Fig. 7g and Supplementary Fig. 12e)[72,74], which might be due to the high expression of these TEs in both normal and T2D alpha cells, emphasizing the importance of analysis at single-cell resolution.

As a final human disease dataset, we reanalyzed a glioblastoma scRNA-seq experiment[75], and were able to identify TEs specifically expressed in neoplastic cells and that were correlated with the expression of *EGFR* (Supplementary Fig. 12f–h), a gene

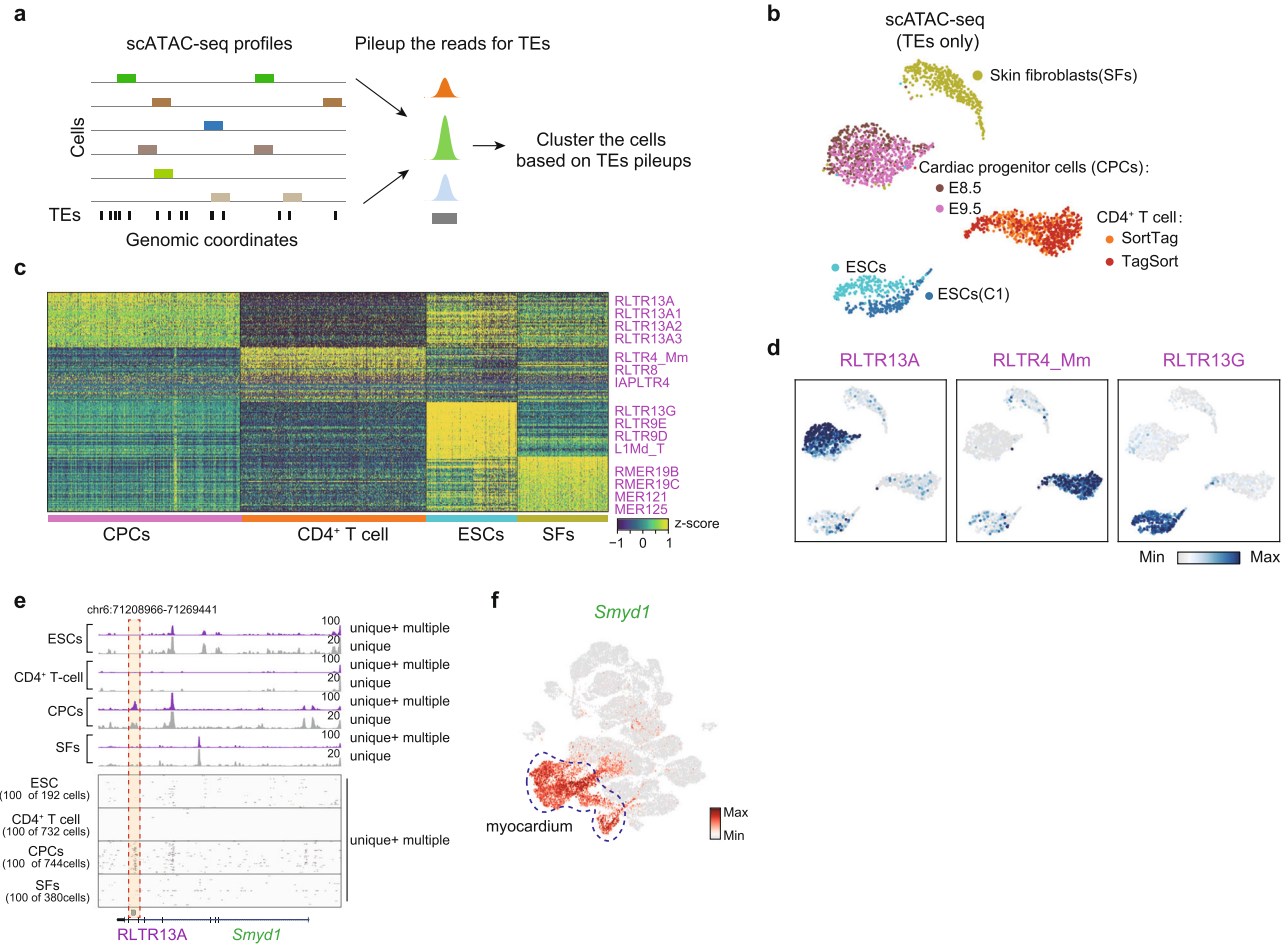

Fig. 6 Analysis of the chromatin state of TEs in single-cell ATAC-seq data. a Schematic plot of scTE for scATAC-seq data analysis. The reads are mapped to the genome, and assigned to a metagene TE, and then the cells were clustered based on the TE matrix. **b** UMAP plot of the TE chromatin state from scATAC-seq data for a selection of FACS-purified mouse cell types. **c** Heatmap of the top 50 cell type-specific opened TEs in the indicated cell types, selected example TEs are indicated. **d** UMAP plot as in (**b**), but cells are colored by chromatin-state of the indicated TEs. **e** Genome tracks showing the aggregate scATAC-seq profiles (top panel). Randomly selected 100 single cell profiles are show below the aggregated profiles (bottom panel). Which include (unique + multiple), or exclude (unique), multiple mapped reads. **f** UMAP plot of the expression of the myocardium marker gene *Smyd1*, from the cardiogenesis data, see Fig. 3i.

upregulated in a large percentage of glioblastomas[75]. Above all, these results revealed the dysregulation of TE expression in diseased human cells, which deserves further mechanistic study and may help to identify new diagnostic markers and therapeutic targets.

Discussion

TEs are the most abundant elements in the genome, however, the understanding of their impact on genome evolution, function and disease remains limited. The rise of genomics and large-scale high-throughput sequencing has shed light on the multi-faceted role of TEs. However, many genomic studies exclude TEs due to difficulties in their analysis as a consequence of their repetitive nature[22]. Thus, TE analysis often requires the use of specialized tools to extract meaning[5,23]. Here, we developed scTE specifically for the analysis of TEs from single-cell sequencing data. By taking advantage of this tool, we could observe previously identified phenomena such as MERVL and LTR7/HERVH expression in mouse and human ESCs, respectively. We then observed widespread heterogeneity of TE expression throughout embryonic development, in mature somatic cells, during the reprogramming process and in human diseases, and discovered a wealth of cell fate-specific TE expression. These associations cannot be observed when only considering bulk samples, demonstrating the power of single-cell sequencing, and the importance of analyzing TE expression. A recent study[76] reported quantification of transposable elements chimeric transcripts in single-cell RNA-seq data assisted by transcript assembly, and identified heterogeneously expressed TE transcripts during mouse gastrulation and early organogenesis, which is consistent and complementary with our findings.

One of the key findings of our analysis has revealed the various TEs that are specifically expressed in different cell types. The expression of TEs during the pre-implantation development stage has been demonstrated previously[7], our findings extend this to gastrulation and early organogenesis. We find a wide array of expression of TEs in the extraembryonic tissues, which may be related to their activity as enhancers[77]. Furthermore, we show the expression of TEs within the specific lineages in the developing fetal heart. In addition, TEs are also heterogeneously expressed between cell types in adult somatic cells, which has not been demonstrated before, as TEs are thought to be primarily silent in adult tissues. Notably, we found a vast of trove of TEs that are expressed in the brain and the immune system, and individual TE types that are specifically expressed in different sub cell types.

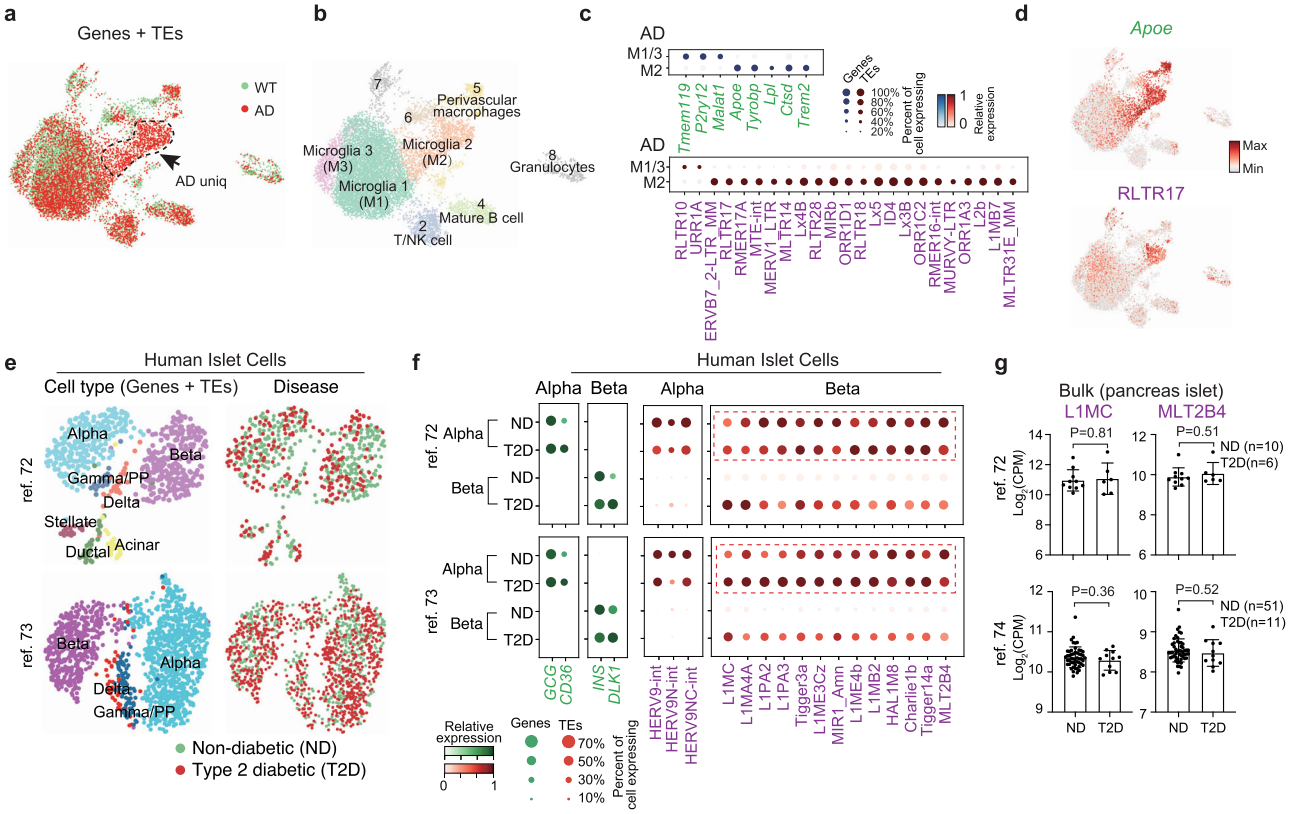

Fig. 7 TEs are differentially expressed in single cells in the diseased state. a UMAP plot of the single cells genes and TE expression, cells are colored by WT (wild-type) and AD (Alzheimer's disease) state. **b** UMAP plot, as in (**a**), but clustered into groups (Leiden, resolution = 0.5). **c** Dot plot showing the differential expressed genes (top) and TEs (bottom) between disease-associated microglia (M2) and homeostatic microglia (M1/3) in AD mice. **d** UMAP plot, as in (**a**), but cells are colored by the expression of the indicated *Apoe* or the TE RLTR17. **e** UMAP plots of pancreatic islet cells. Cells are colored by cell types (left) or disease-state (right). Cell types were annotated according to the metadata from the original study, and matched the expression of known marker genes. **f** Dot plot showing marker gene expression (green) or TEs (red) differentially expressed between healthy and T2D alpha and beta cells (Benjamini–Hochberg corrected Wilcoxon rank-sum test, $P < 0.01$, and at least >2-fold change between groups). **g** Bar charts showing the expression of the indicated TEs from bulk RNA-seq data. Data are presented as mean values ± SD. *P*-value was from an unpaired Student's *t*-test.

Considering the close relationship between the evolution of immune system, brain and TEs[47–49], these results hint at further functions for TEs in these two systems.

How cells decide their fate is a fundamental question in biology. Stem cell differentiation and somatic cell reprogramming are both powerful in vitro models that mimic in vivo development and have provided great insight into cell fate decisions. However, how TEs are involved in these processes is still largely unknown. In this study, we have identified the TEs LTR32 and MLT1H1 that were differentially regulated between contractile and non-contractile cell fate decisions during human cardiac differentiation. In addition, we also observed the divergent expression of ERVB7_1-LTR_Mm and IAP elements during reprogramming. Whereas ERVB7_1-LTR_Mm was highly expressed in iPSCs, IAP elements were silenced at the final stage, just before commitment to iPSCs formation (Fig. 5b, f, h). These mechanisms are shared among the Yamanaka factor based and chemical based reprogramming systems, indicating a tight association between TE expression and cell fate.

Overall, whilst the information content of TEs is lower than that of genes, TEs are a useful addendum to the gene information, and, in some cases, they are a major source of information on their own. For example, MERVL expression alone is capable of discriminating 2C-like cells. The routine inclusion of TEs in scRNA-seq analysis pipelines will identify more instances like the 2C/MERVL relationship, and enrich our understanding of cell type, diseases and TE expression control. In addition to scRNA-seq, TE information may be particularly informative in scATAC-seq, and other scChIP-seq-like data. As scATAC-seq is so sparse, individual peaks in individual cells are challenging to resolve. However, by merging TE data it may be possible to infer TEs as enhancer information in single cells.

Considering the growing implication that TEs are important contributors to human disease, their study is becoming increasingly important. In addition to the ability of TEs to impact genomic stability as they duplicate[78], which has clear implications for the development of cancer[79], TEs are also playing more subtle roles in epigenetic control and transcript expression. For example, TEs are spliced into chimeric transcripts that drive the expression of oncogenes[12]. Similarly, the expression of TEs has been associated with several nervous system-related disorders, including neurodegeneration[11], and L1 LINE expression is important in inflammation during aging[80]. In our work, we demonstrate that in single cells of the pancreas there is substantial TE expression deregulation in the beta cells, which is suggestive of epigenetic dysfunction and a loss of control over TE expression. Critically, this observation cannot be observed from bulk pancreatic islet samples. Considering the growing importance of exploring human disease using primary patient samples, the analysis of TEs should be included. However, to date the contribution of TE expression to the aging and diseased states remains relatively unexplored. Our approach will be an important tool in understanding the contributions of TEs to cellular heterogeneity in a variety of systems and in human disease.

Methods

Software availability. scTE is available at https://github.com/JiekaiLab/scTE. The code is freely available and is released under the MIT license. scTE requires Python >3.6, and the python module numpy, scTE supports the Linux and Mac platforms. Software code for the analysis of the data in this paper can be found at: https://github.com/JiekaiLab/scTE/tree/master/example.

scTE pipeline. The input data for scTE consists of the annotation files for genes and TEs, and alignment files in either the SAM or BAM format[81]. By default, scTE uses GENCODE[82] and the UCSC genome browser Repeatmasker track[83] annotations for genes and TEs, respectively. The SAM/BAM file contains the aligned read genome locations. Many alignment programs can distinguish reads that have a unique alignment in the genome (termed unique-reads) or map to multiple genomic loci (termed multimapping reads or non-unique reads). Multimapping reads are critical for TE quantification, as TEs contain many repeated sequences and non-unique reads often map inside the TEs. To get an accurate quantitation of the number of reads mapping to TEs these reads should be preserved. However, in many analyses pipelines these reads are discarded. scTE recommends aligners to keep all of the mapped reads, and we recommend that only the best single aligned multimapped read be kept. The reads can be aligned by any genome aligner, but the aligned reads must be against the genome (i.e., not against a set of genes or transcript assembly). scTE is most tuned to STAR-solo[27] or the Cell Ranger pipeline outputs, and can accept BAM files produced by either of these two programs. For other aligners, the barcode should be stored in the 'CR:Z' tag, and the UMI in the 'UR:Z' tag in the BAM file. If the UMI is missing or not used in the scRNA-seq technology (for example, on the Fluidigm C1 platform), it can be disabled with –UMI False (the default is True) switch in scTE. If the barcode is missing it can be disabled with the –CB False (the default is True), and instead the cell barcodes will be taken from the names of the BAM files (multiple BAM files can be provided to scTE with the –i option).

scTE gene and TE indices. scTE builds genome indices for the fast alignment of reads to genes and TEs. These indices can be automatically generated using the commands:

scTE_build -g mm10 # mouse genome,
scTE_build -g hg38 # human genome.

These two scripts will automatically download the genome annotations, for mouse:
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M21/gencode.vM21.annotation.gtf.gz,
http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/rmsk.txt.gz.
Or for human:
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_30/gencode.v30.annotation.gtf.gz,
http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz.

These annotations are then processed and converted into genome indices. The scTE algorithm will allocate reads first to gene exons, and then to TEs, by default. Hence TEs inside exon/UTR regions of genes annotated in GENCODE will only contribute to the gene, and not to the TE score. This feature can be changed by setting '–mode/-m exclusive' in scTE, which will instruct scTE to assign the reads to both TEs and genes if a read comes from a TE inside exon/UTR regions of genes.

Analysis of 10x-style data. scRNA-seq data was processed using the scTE 10x pipeline, Briefly, reads were aligned to the genome using STARsolo[27] with the setting '--outSAMattributes NH HI AS nM CR CY UR UY --readFilesCommand zcat --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --runRNGseed 777 --outSAMmultNmax 1'. The default scTE parameters for 10x were used to get the molecule count matrix. The count matrix was lightly filtered to exclude cell barcodes with low numbers of counts: Cells with less than 1000 UMIs and less than 500 genes detected were filtered out, and only the top 10,000 cells with the highest gene count were kept (these default setting can be altered with the '--expect-cells, --min_count and --min_genes' switches in scTE, note that the cell counts are further filtered on a case-by-case basis for each experiment, as detailed below). Other downstream analysis was performed by SCANPY[25]. Specific analysis settings for the individual datasets are described below. Normalized expression, used in the UMAP plots, is calculated using the normalize_total function in scanpy, or the calculateSumFactors from SCRAN which estimates size factors for each cell to remove bias within the cell counts, and improve cross-cell comparison of cell expression values. Relative expression values, used in the dotplots and heatmaps scales the expression within the range 0 to 1, representing the minimum or maximum relative expression across a set of cells or clusters.

Analysis of C1/SMART-seq-style data. scRNA-seq data were processed using the scTE C1/SMART–seq pipeline, Briefly, reads were aligned to the genome using STAR[27], with the setting '--winAnchorMultimapNmax 100 --outSAMmultNmax 1 --outSAMmultNmax 1'. The default scTE parameters for C1/SMART-seq were used to get the molecule count matrix. Cells with less than 10,000 counts and less than 2000 expressed genes were filtered out. Cells with more than 20% fraction of mitochondrial counts were discarded. Downstream analysis was performed the same as for the 10x data pipeline. Fluidigm C1/SMART-seq data comes as a single BAM file per barcode. To analyze this data, the 'barcode' is taken from the input BAM filenames, and both -CB and -UMI should be False:

scTE -i *.bam -p 4 -o <output_name> --genome mm10 -x mm10.exclusive.idx -CB False -UMI False.

The resulting matrices can then be integrated into an scRNA-seq analysis pipeline.

Analysis of human cardiac differentiation scRNA-seq data. The raw data were download from E-MTAB-6268[35]. As this data were generated using the Single Cell 3′ Library, Gel Bead and Multiplex kit (version 1, 10x Genomics, Cat. #PN-120233), the cell barcode and UMI sequence are not in the same read. First, we merged the cell barcode and UMI sequence into the same read using a custom script, and then aligned the modified fastq file to the hg38 genome using STARsolo, as described above. Cells with less than 500 expressed genes/TEs and cells that have more than 20% fraction of mitochondrial reads were discarded. Single cell trajectory was analyzed by Harmony[84] and the top 1000 highly variable genes were used for PCA, and the force directed layout was computed using first 150 PCs (principle components). Differentially expressed genes and TEs were analyzed using the SCANPY rank_genes_groups functions by t-test method, the top 500 specifically expressed TEs and genes with Benjamini–Hochberg corrected p-value <0.01 and log2(fold-change) >0.5 are selected for downstream analysis.

Analysis of the gastrulation scRNA-seq data. The raw data were download from E-MTAB-6967, and aligned to the mm10 genome using STARsolo[27], with the parameters "--readFilesCommand zcat --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --runRNGseed 777 --outSAMmultNmax 1". Cells with less than 3000 expressed genes/TEs, and less than 8000 UMIs were discarded. Genes expressed in less than 50 cells were removed from the analysis. The count matrix was normalized using normalize_total function of SCANPY, and the top 2000 most highly variable genes were used for PCA, and the first 20 PCs (principle components) were used, as described in the original publication[17]. UMAP plots were generated (min_dist=0.6). Data is from E-MTAB-6967[17].

Analysis of *Tabula Muris* scRNA-seq data. The C1/Smart-seq2 scRNA-seq raw data was download from GSE109774[43], the reads were aligned to the mm10 genome using STAR with the parameters '--readFilesCommand zcat --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --runRNGseed 777 --outSAMmultNmax 1'. The genes/TEs and cell expression matrix was generated using scTE. Cells with less than 50000 counts or more than 2^7 counts, less than 1000 expressed genes, or more than 20% fraction of mitochondrial counts were removed. The filtered matrix was normalized using scran[85]. The top 4000 most highly variable genes were used for PCA, and the first 50 PCs were used for downstream analysis. The cell cluster specific expressed genes/TEs was calculated using SCANPY rank_genes_groups functions by t-test method, the top 500 specifically expressed TEs and genes with Benjamini–Hochberg corrected p-value <0.01 and log2(fold-change) >0.5 compare to all other groups of cells were kept.

Analysis of the OKSM/chemical reprogramming data. The raw data were download from GSE115943[54] and GSE114952[31]. Cells with less than 10000 UMIs or more than 1,000,000 UMIs, or expressed less than 1000 expressed genes, or more than 20% fraction of mitochondrial counts were removed. The filtered matrices were normalized using scran[85]. The top 4000 most highly variable genes were used for PCA, and the first 50 PCs were used for downstream analysis. The cell trajectory routes were taken from the original studies. Differentially expressed genes/TEs were calculated using SCANPY rank_genes_groups functions by the t-test method, the TEs and genes with Benjamini–Hochberg corrected p-value <0.01 and log2(fold-change) >0.5 compared to all other branches of cells were kept.

Analysis of the OKS reprogramming data. The C1/SMART-seq data were taken from GSE103221[28]. the reads were aligned to the mm10 genome using STAR with the parameters "--readFilesCommand zcat --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --runRNGseed 777 --outSAMmultNmax 1". The genes/TEs and cells expression matrix was generated using scTE. Cells with less than 10,000 counts or more than 2^7 counts, less than 1000 expressed genes, or more than 20% fraction of mitochondrial counts were removed. The filtered matrix was normalized using scran[85]. The top 4000 most highly variable genes were used for PCA, and the first 50 PCs were used for downstream analysis. The genes/TEs expression trajectories on pseudotemporal orderings of cells (Fig. 5e) were analyzed by LineagePulse (https://github.com/YosefLab/LineagePulse) according to the pseudotime taken from the original study.

Analysis of the embryonic heart scRNA-seq data. The raw data was download from GSE126128[42]. This data was aligned to the genome using STARsolo[27], as described above. Cells with less than 3000 expressed genes/TEs and the cells with

less than 8000 UMIs or more than 100,000 UMIS were deleted from the analysis. The count matrix was normalized using normalize_total function of SCANPY. The top 2000 most highly variable genes were used for PCA, and the first 20 PCs were used for downstream analysis. UMA projections were generated (min_dist=0.7).

Analysis of Alzheimer's disease scRNA-seq data. The MARS-seq scRNA-seq raw data were download from GSE98969[71]. The raw fastq file were modified using custom scripts to embed the cell barcode and UMI in the same read, as in the 10x scRNA-seq format. The modified reads were aligned to the mm10 genome with STARsolo as described above. Cells with less than 5000 UMIs or more than 1,000,000 UMIs, or expressed less than 500 genes, or more than 20% fraction of mitochondrial counts, were removed. The filtered matrix was normalized using scran[85]. The top 4000 most highly variable genes were used for PCA, and the first 50 PCs were used for downstream analysis. The differentially expressed genes and TEs between M2 and M1/3 were analyzed using SCANPY rank_genes_groups functions by t-test method, the genes or TEs with Benjamini–Hochberg corrected *p*-value <0.01 and log2(fold-change) >0.5 compared to each other were kept.

Analysis of the type 2 diabetes/glioblastoma sc-RNA-seq data. The raw data was download from GSE86473[72], GSE81608[73]. The data was aligned to the hg38 genome using STAR[27], as described above for C1 data. Cells with less than 5000 expressed genes/TEs and cells with less than 1×10^6 counts or more than 6×10^6 or were deleted from the analysis. The count matrix was normalized using the normalize_total function of SCANPY. There was a strong batch effect based on the sex of the donor in the type 2 diabetes datasets, this was removed using the regress_out function of SCANPY[25]. We did not detect any other batch effect from other confounding variables (age, body-mass index, race). The top 2000 most highly variable genes were used for PCA, and the first 15 PCs (type 2 diabetes) or 25 PCs (glioblastoma) were used. UMAP plots were generated using SCANPY (min_dist = 0.7).

Bulk RNA-seq analysis. Analysis of bulk RNA-seq data was performed essentially as previously described[3,86], with some modifications. Briefly, reads were aligned to the mouse or human genome/transcriptome (GENCODE transcript annotations, mouse M21 or human 30) using STAR (v2.7.1a)[27]. TEtranscripts[87] or scTE (with the setting -CB False -UMI False) was used to quantitate reads on TEs. Reads were GC normalized using EDASeq (v2.16.3)[88], and analyzed using glbase[89].

Motif enrichment analysis. The TF motif enrichment in TEs (Supplementary Fig. 6c and 9k) was measured using AME from the MEME suite[90] with the options "--control --shuffle".

Bulk ATAC-seq analysis. Analysis of bulk ATAC-seq data was performed essentially as previously described[3,91]. Briefly, reads were aligned to the mouse or human genome (mm10 or hg38) using bowtie2[92] (v2.3.5.1), with the options: "-p 6 --mm --very-sensitive --no-unal --no-mixed --no-discordant -X2000", and reads mapping to TEs were counted using te_counter (https://github.com/oaxiom/te_counter). The counts per million (CPM) reads metric was used for enrichment scores.

ChIP-seq data analysis. Analysis of ChIP-seq data was performed as previously described[3]. Briefly, reads were mapped to mouse genome (mm10) genome using bowtie2[92] with the options: -p 20 --very-sensitive --end-to-end --no-unal. For pair-end sequence data, only concordantly aligned pairs were kept. All mapped reads were kept, but only the best alignment is reported for multimapped reads, if more than one equivalent best alignment was found, then one random alignment was reported. Alignment bam files were transformed into read coverage files (bigwig format) using deepTools[93] with the RPKM (reads per kilobase per million mapped reads) normalization method.

Analysis of the scATAC-seq data. Three datasets were used for to test scTE performance on scATAC-seq data. Presorted mouse cells and unsorted mouse spleen cells, using a custom scATAC-seq technology[64], and human PBMC data from 10xgenomics. The first two datasets could be aligned directly to the mouse/human genome. The 10xgenomics data required preprocessing: we downloaded the scATAC-seq data from the 10xgenomics website (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_pbmc_10k_v1). The barcode was inserted into the read name, so that the mapping could keep track of the cell ID. This yielded read names inside the FASTQ, such as: (where CCACGTTGTGGACTGA sequence is the cell barcode).

@CCACGTTGTGGACTGA:A00519:269:H7FM2DRXX:1:1101:1325:1000 1:N:0:AAGCATAA.

The genome indices were prebuilt using:

wget -c -O mm10.te.txt.gz 'http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/rmsk.txt.gz',

zcat mm10.te.txt.gz | grep -E 'LINE|SINE|LTR|Retroposon|DNA'|cut -f6-8,11>mm10.te.bed,

python3/share/apps/genomics/unstable/scTE/bin/scTEATAC_build -g mm10.te.bed -o mm10.te.atac,

wget -c -O hg38.te.txt.gz 'http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz',

zcat hg38.te.txt.gz | grep -E 'LINE|SINE|LTR|Retroposon|DNA'|cut -f6-8,11>hg38.te.bed,

python3/share/apps/genomics/unstable/scTE/bin/scTEATAC_build -g hg38.te.bed -o hg38.te.atac.

The data were aligned to the mouse mm10 or human hg38 genome using bowtie2[92] with the command options "-p 6 --mm --very-sensitive --no-unal --no-mixed --no-discordant -X2000". The resulting data was then processed using scTE with the command:

scTEATAC -i <in> -x <genome>.te.atac.idx -g <genome> -p 1 -UMI False -CB True -o <out>.

scTE will internally deduplicate reads, by allowing only a single read per base pair of the genome. scTE will produce a matrix containing cell barcodes (rows) and TEs (columns). The information across all genomic TEs is merged into a single TE subtype. This matrix is then processed in a manner similar to RNA-seq. TEs were first filtered to remove low "expressed" TEs with less than 1000 read counts, then samples were normalized using SCANPY or scran, and TE counts placed onto a normalized scale. Downstream analysis used SCANPY.

Quantitative PCR. Total RNAs were extracted by chloroform-isopropanol method. The first-strand cDNAs were synthesized with ReverTra Ace (Toyobo) and oligo-dT (Takara), and then qRT-PCR was performed on a CFX96 real-time system (Bio-Rad) with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad). The primers used for qRT-PCR were listed in Supplementary Table 2.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing datasets used in this study were obtained from public data repositories. Detailed information, including accession URLs for published datasets are available in Supplementary Table 1. All relevant data are available from the corresponding authors on reasonable request.

Code availability

The full package of scTE[94] is available at: https://github.com/JiekaiLab/scTE.

Received: 28 May 2020; Accepted: 4 February 2021;
Published online: 05 March 2021

References

1. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
2. 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.* **60**, 1722–1733 (2015).
3. He, J. et al. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat. Commun.* **10**, 34 (2019).
4. Lu, X. et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423–425 (2014).
5. Bourque, G. et al. Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
6. Wei, L. & Cao, X. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Sci. China Life Sci.* **59**, 24–37 (2016).
7. Goke, J. et al. Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135–141 (2015).
8. Grow, E. J. et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
9. Percharde, M. et al. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391–405.e319 (2018).
10. Jachowicz, J. W. et al. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet* **49**, 1502–1510 (2017).
11. Tam, O. H., Ostrow, L. W. & Gale Hammell, M. Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. *Mob. DNA* **10**, 32 (2019).
12. Jang, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).
13. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

14. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
15. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet* **42**, 631–634 (2010).
16. Faulkner, G. J. et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet* **41**, 563–571 (2009).
17. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
18. Li, H. et al. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775–789 e718 (2019).
19. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
20. Macfarlan, T. S. et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
21. Fu, X., Wu, X., Djekidel, M. N. & Zhang, Y. Myc and Dnmt1 impede the pluripotent to totipotent state transition in embryonic stem cells. *Nat. Cell Biol.* **21**, 835–844 (2019).
22. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).
23. Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nat. Rev. Genet.* **19**, 688–704 (2018).
24. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902 e1821 (2019).
25. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
26. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
27. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
28. Guo, L. et al. Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Mol. Cell* **73**, 815–829 e817 (2019).
29. Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101–4114 (2016).
30. Rodriguez-Terrones, D. & Torres-Padilla, M. E. Nimble and ready to mingle: transposon outbursts of early development. *Trends Genet.* **34**, 806–820 (2018).
31. Zhao, T. et al. Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell Stem Cell* **23**, 31–45 e37 (2018).
32. Wang, J. et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
33. Fort, A. et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46**, 558–566 (2014).
34. Theunissen, T. W. et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* **19**, 502–515 (2016).
35. Friedman, C. E. et al. Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. *Cell Stem Cell* **23**, 586–598 e588 (2018).
36. Liu, Q. et al. Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. *Circ. Res.* **121**, 376–391 (2017).
37. Abed, M. et al. The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. *PLoS ONE* **14**, e0214110 (2019).
38. Zhong, Y. et al. Isolation of primitive mouse extraembryonic endoderm (pXEN) stem cell lines. *Stem Cell Res.* **30**, 100–112 (2018).
39. Factor, D. C. et al. Epigenomic comparison reveals activation of "seed" enhancers during transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 854–863 (2014).
40. Morey, L. et al. Polycomb regulates mesoderm cell fate-specification in embryonic stem cells through activation and repression mechanisms. *Cell Stem Cell* **17**, 300–315 (2015).
41. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599 (2012).
42. de Soysa, T. Y. et al. Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature* **572**, 120–124 (2019).
43. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
44. Wu, J. Q. et al. Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS Genet.* **8**, e1002565 (2012).
45. Kim, K. Y. et al. Uhrf1 regulates active transcriptional marks at bivalent domains in pluripotent stem cells through Setd1a. *Nat. Commun.* **9**, 2583 (2018).
46. Adachi, K. et al. Context-dependent wiring of Sox2 regulatory networks for self-renewal of embryonic and trophoblast stem cells. *Mol. Cell* **52**, 380–392 (2013).
47. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
48. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet* **16**, 184–192 (2015).
49. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* **15**, 497–506 (2014).
50. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
51. Wang, B. et al. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. *Cell Rep.* **27**, 3473–3485 e3475 (2019).
52. Hou, P. et al. Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* **341**, 651–654 (2013).
53. Cao, S. et al. Chromatin accessibility dynamics during chemical induction of pluripotency. *Cell Stem Cell* **22**, 529–542 e525 (2018).
54. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 e922 (2019).
55. Friedli, M. et al. Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Res.* **24**, 1251–1259 (2014).
56. Liu, J. et al. The oncogene c-Jun impedes somatic cell reprogramming. *Nat. Cell Biol.* **17**, 856–867 (2015).
57. Chronis, C. et al. Cooperative binding of transcription factors orchestrates reprogramming. *Cell* **168**, 442–459.e420 (2017).
58. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression. *Nat. Rev. Genet* **21**, 721–736 (2020).
59. O'Neill, K., Brocks, D. & Hammell, M. G. Mobile genomics: tools and techniques for tackling transposons. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190345 (2020).
60. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
61. Shema, E., Bernstein, B. E. & Buenrostro, J. D. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet* **51**, 19–25 (2019).
62. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e1318 (2018).
63. Sun, X. et al. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc. Natl Acad. Sci. USA* **115**, E5526–E5535 (2018).
64. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).
65. Warren, J. S. et al. Histone methyltransferase Smyd1 regulates mitochondrial energetics in the heart. *Proc. Natl Acad. Sci. USA* **115**, E7871–E7880 (2018).
66. Rasmussen, T. L. et al. Smyd1 facilitates heart development by antagonizing oxidative and ER stress responses. *PLoS ONE* **10**, e0121765 (2015).
67. Franklin, S. et al. The chromatin-binding protein Smyd1 restricts adult mammalian heart growth. *Am. J. Physiol. Heart Circ. Physiol.* **311**, H1234–H1247 (2016).
68. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet* https://doi.org/10.1038/s41588-019-0505-9 (2019).
69. Bravo Gonzalez-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
70. Li, W. et al. Activation of transposable elements during aging and neuronal decline in *Drosophila*. *Nat. Neurosci.* **16**, 529–531 (2013).
71. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290 e1217 (2017).
72. Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* **27**, 208–222 (2017).
73. Xin, Y. et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* **24**, 608–615 (2016).
74. Fadista, J. et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl Acad. Sci. USA* **111**, 13924–13929 (2014).
75. Darmanis, S. et al. Single-cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
76. Shao, W. & Wang, T. Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. *Genome Res* **31**, 88–100 (2021).
77. Chuong, E. B., Rumi, M. A., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet* **45**, 325–329 (2013).

78. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nat. Rev. Genet* https://doi.org/10.1038/s41576-019-0165-8 (2019).
79. Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
80. De Cecco, M. et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566**, 73–78 (2019).
81. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
83. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
84. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
85. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
86. Hutchins, A. P. et al. Models of global gene expression define major domains of cell type and tissue identity. *Nucleic Acids Res.* **45**, 2354–2367 (2017).
87. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
88. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinforma.* **12**, 480 (2011).
89. Hutchins, A. P., Jauch, R., Dyla, M. & Miranda-Saavedra, D. glbase: a framework for combining, analyzing and displaying heterogeneous genomic and high-throughput sequencing data. *Cell Regen. (Lond.)* **3**, 1 (2014).
90. McLeay, R. C. & Bailey, T. L. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinforma.* **11**, 165 (2010).
91. Li, D. et al. Chromatin accessibility dynamics during iPSC reprogramming. *Cell Stem Cell* **21**, 819–833 e816 (2017).
92. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
93. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
94. He, J. et al. Identifying transposable element expression dynamics and heterogeneity during development at the single cell level with a processing pipeline scTE, https://github.com/JiekaiLab/scTE, https://doi.org/10.5281/zenodo.4420937 (2021).

Acknowledgements

We are grateful to Rujin Huang for the discussions and constructive suggestions. We appreciate the assistance of Kaixin Wu, Lihui Lin, Huijian Feng, and Yuanbang Mai on the data analysis and qPCR experiment. We thank the Guangzhou Branch of the Supercomputing Center of Chinese Academy of Sciences, the Center for Computational Science and Engineering of Southern University of Science and Technology, and the Cloud Computing Center of Chinese Academy of Sciences for their support. This work was supported by the National Key R&D Program of China (2019YFA0110200), the Frontier Science Research Program of the CAS (ZDBS-LY-SM007), the Key Research and Development Program of Guangzhou Regenerative Medicine and Health Guangdong Laboratory (2018GZR110104003, 2019GZR110108001), the National Natural Science Foundation of China (31970589, 31801217, 31850410463, 31850410486), Science and Technology Planning Project of Guangdong Province, China (2020B1212060052), The Science and Technology Program of Guangzhou (201804020052), Guangdong Science and Technology Commission (2019A050510004), and the Shenzhen Peacock plan (201701090668B).

Author contributions

J.C. conceived the project. J.H., A.P.H., and J.C. initiated the project and wrote the paper. J.H. and A.P.H. performed the bioinformatic analysis with the assistance of all other authors. A.P.H. and J.C. supervised and funded the project.

Competing interests

The authors declare no competing interests

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21808-x.

Correspondence and requests for materials should be addressed to A.P.H. or J.C.

Peer review information *Nature Communications* thanks Molly Hammell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.