MATES: A Deep Learning-Based Model for Locus-specific Quantification of Transposable Elements in Single Cell

Ruohan Wang[1,2†], Yumin Zheng[2,3†], Zijian Zhang[4†], Xiaopeng Zhu[5], Tao P. Wu[4*], Jun Ding[1,2,3,6*]

[1]School of Computer Science, McGill University, Montreal, Quebec,Canada.
[2]Meakins-Christie Laboratories, Department of Medicine, McGill University, Montreal, Quebec,Canada.
[3]Quantitative Life Sciences, Faculty of Medicine & Health Sciences, McGill University, Montreal, Quebec,Canada.
[4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA.
[5]MyCellome LLC., Pittsburgh, PA,USA.
[6]Mila-Quebec AI Institue, Montreal, Quebec, Canada.

*Corresponding author(s). E-mail(s): tao.wu@bcm.edu; jun.ding@mcgill.ca;
†These authors contributed equally to this work.

Abstract

Transposable elements (TEs) are crucial for genetic diversity and gene regulation. Current single-cell quantification methods often align multi-mapping reads to either 'best-mapped' or 'random-mapped' locations and categorize them at sub-family levels, overlooking the biological necessity for accurate, locus-specific TE quantification. Moreover, these existing methods are primarily designed for and focused on transcriptomics data, which restricts their adaptability to single-cell data of other modalities. To address these challenges, here we introduce MATES, a novel deep-learning approach that accurately allocates multi-mapping reads to specific loci of TEs, utilizing context from adjacent read alignments flanking the TE locus. When applied to diverse single-cell omics datasets, MATES shows improved performance over existing methods, enhancing the accuracy of TE quantification and aiding in the identification of marker TEs for identified cell populations. This development enables exploring single-cell heterogeneity and gene regulation through the lens of TEs, offering a transformative tool for the single-cell genomics community.

Keywords: Transposable elements, Multi-mapping reads, Deep neural network, Transposon quantification, Cell clustering, Transposon biomarkers

Introduction

Transposable elements (TEs), also known as transposons or jumping genes, form a significant part of mammalian genomes and play crucial roles in gene regulation, genome evolution, and cell-to-cell heterogeneity [1–3]. However, the understanding of TEs at the single-cell level has been limited due to challenges in quantifying multi-mapping sequencing reads, a result of their repetitive sequences and high copy numbers [4]. Studying TEs in single-cell genomics, in contrast to bulk sequencing, is crucial for understanding their dynamic regulation and contribution to cellular heterogeneity. This approach reveals the complex expression patterns of TEs and their significant impact on the transcriptional landscape [5, 6]. The variability in TE activity across individual cells contributes to the complexity of gene regulation and cellular dynamics, crucial in both normal development and disease states like

cancer [7, 8]. Advancing our understanding of TEs in single-cell genomics not only promises to unveil the intricacies of cellular processes and evolutionary biology but also poses significant challenges due to the repetitive nature of TE sequences. Developing refined quantification and analytical methods is essential to fully harness the potential of single-cell TE analysis, potentially leading to breakthroughs in comprehensive understanding cellular states and its implications in health and disease.

While there is an abundance of methods for TE quantification at the bulk level, such as Squire [9] and Telescope [10], emerging single-cell methods like scTE [2] and SoloTE [11] remain limited. SoloTE addresses the challenge of repetitive sequence assignment by allocating multi-mapping reads to a singular locus based on the best read alignment score. In contrast, scTE employs a similar strategy but rectifies multi-mapping errors using the count of unique-mapping reads. However, these methods often rely heavily on alignment algorithms to address multi-mapping reads, neglecting the genomics context flanking the TEs. Furthermore, current methods fail to provide an accurate locus-specific TE quantification. scTE measures TE expression by allocating reads to corresponding TE subfamilies. SoloTE did present locus-level TE expression, but it only maps the reads into the "best" location; thus, it is very limited in handling multi-mapping reads that prevail in TE regions[12, 13]. These approaches overlook or circumvent the challenging multi-mapping read assignment rooted in the repetitive characteristics of TEs. This oversight can result in an underestimation of the complexity and uncertainty associated with assigning multi-mapping reads in TE quantification. The frequent occurrence of multi-mapping reads across TE regions highlights the existing gaps in single-cell TE quantification techniques, emphasizing the need for precise read assignment to specific genomics loci. An accurate quantification of locus-specific TE expression is paramount to understanding the mechanisms of TEs' crosstalk to regional genome regulations[5, 14].

Recent advancements in single-cell sequencing technologies have broadened the scope to include a range of modalities [3, 15, 16], extending beyond traditional transcriptomics. These developments allow for the analysis of different cellular components, such as the epigenome, in addition to the transcriptome, within individual cells. Furthering this progression, single-cell multi-omics methods, like 10x Genomics' Multiome, can even simultaneously profile both the transcriptome and epigenome in the same cells. This dual-modality profiling is highlighted in comprehensive reviews of the technological landscape and applications in single-cell multi-omics, demonstrating a significant impact on molecular cell biology and the potential to decipher complex biological processes at the single-cell level [17–20]. Despite these advancements, current TE quantification methods, primarily designed for single-cell transcriptomics, face limitations when dealing with data from other modalities, such as single-cell assays for transposase-accessible chromatin sequencing (scATAC-seq). These methods also lack comprehensive solutions for the combined quantification and analysis of TEs in multi-omics datasets. Consequently, there is a pressing need for methodologies capable of accurately aligning multi-mapping TE reads and performing locus-level quantification across various modalities, underscoring the ongoing challenges and opportunities in the field of single-cell genomics and multi-omics research, particularly through the evolving lens of TEs.

To address these challenges and fill the gap, here we introduce MATES (Multi-mapping Alignment for TE loci quantification in Single-cell), a novel deep neural network-based method tailored for accurate locus-specific TE quantification in single-cell sequencing data across modalities. MATES harnesses the distribution of uniquely mapped reads occurrence flanking TE loci and assigns multiple mapping TE reads for locus-specific TE quantification. MATES leverages the power of the deep neural network to capture complex relationships between the context distribution of unique-mapping reads flanking TE loci and the probability of multi-mapping reads assigned to those loci, allowing it to handle the multi-mapping read assignments probabilistically based on the local context background of the TE loci.

In our systematic evaluation of MATES using various single-cell datasets across different modalities and sequencing platforms, we demonstrated that MATES consistently outperforms prevalent TE quantification methods. Beyond achieving higher precision in TE quantification, the capability of MATES for locus-specific TE quantification and its ability to generalize into single-cell data across various sequencing platforms and data modalities enables a deeper comprehension of how TEs contribute to cellular dynamics and gene regulation in biological processes. To verify the accuracy of our method for quantifying multi-mapping TE reads, we further utilized a paired single-cell dataset of skin tissue in the tumor stage of Melanoma brain metastasis comprising both nanopore long-read and short-read 10x sequencing for the same set of cells. The nanopore long-read sequencing can serve as the ground truth as the length of the long-read significantly reduces the occurrence of the multi-mapping reads. By comparing the single-cell TE quantification produced by MATES with the actual

TE quantification empowered by the nanopore long-read sequencing, we have further showcased the accuracy of the tool and its advantage over current methods. Our results demonstrate the effectiveness of MATES in exploring the roles of TEs in single-cell biology, offering a reliable and versatile solution for TE quantification in a variety of experimental contexts. MATES fills the existing gap in single-cell TE quantification techniques, providing a robust and adaptable method that significantly advances our ability to unravel the complex contributions of TEs to cellular heterogeneity and genome regulation. With the burgeoning field of single-cell sequencing data, the potential for in-depth TE quantification and analysis is enormous. This opens avenues to gain invaluable insights into the molecular mechanisms underpinning various biological processes [8].

Results

Methods overview

MATES is a specialized tool designed for precise quantification of TEs in single-cell datasets of different modalities. The method involves several key steps. First, raw reads are mapped to the reference genome, identifying reads that map uniquely to a TE locus (unique reads) and the reads that get mapped into multiple TE loci (multi-mapping reads) (Fig.1 a). Next, we compute the coverage vector for each TE locus, representing the distribution of unique reads surrounding the locus. Each TE region (locus) is then subdivided into smaller bins of length W (e.g., 10 base pairs). Each of these bins is then classified into Unique-dominant (U) or Multi-dominant regions (M) based on the percentage of unique reads and multiple-mapping reads within the bin (Fig.1 b). Third, we employed an AutoEncoder (AE) model to learn the latent embeddings (V_u) to represent the high-dimensional unique reads coverage vector of a TE locus, which represents the reads mapping context flanking the specific TE locus. The one-hot encoded TE family information (T_i) was also taken as input for the model. Fourth, the learned latent embedding (V_u) and the TE family embedding (T_i) will then be used to predict the multi-mapping ratio (α) for the specific TE locus via a multiple-layer perceptron regressor. The grand loss we used to learn the model is composed of two components (L_1 and L_2). The former represents the reconstruction loss of the AutoEncoder while the latter reflects the continuity of the actual reads coverage across neighboring small bins on the TE. In other words, the final read coverage on the multi-mapping dominant (M) bins should be close to its neighboring unique dominant (U) bins because of their genomics proximity. Finally, once we train the model that predicts the multi-mapping ratio for each TE locus, we can then leverage it to count the total number of reads that fall into the specific TE loci, which presents probabilistic quantification of TEs at the locus level (Fig.1 c). By combining TE quantification and conventional gene quantification (e.g., gene expression or gene accessibility) from the single-cell data, which we refer to as "Gene+TE expression" in the sections below, we can then more accurately cluster the cells and identify comprehensive biomarkers (genes and TEs) to characterize the obtained cell clusters (cell subpopulations). Equipped with advanced features, MATES efficiently handles various single-cell data modalities. Its application offers insights into the roles of TEs in diverse datasets, clustering cells, and identifying potential biomarker TEs (Fig.1 d). Beyond its analytical capabilities, MATES also presents locus-specific TE visualization and interpretation. The tool facilitates the generation of comprehensive bigwig files and Interactive Genomic Viewer (IGV) plots, enabling researchers to visually explore and interpret read assignments for TE loci across the genome (Fig.1 e). This capability unlocks the investigation of potential interactions between Transposable Elements (TEs) and the genes situated near TE loci, significantly enhancing our understanding of TE dynamics and their impact on gene regulation and cellular functions.

MATES identifies signature TEs and their specific loci in 2C-like cells (2CLCs) within 10x single-cell RNA-seq data of chemical reprogramming

To demonstrate the precision of MATES in transposable element (TE) quantification from single-cell RNA-seq data, we applied it to a 10x single-cell chemical reprogramming dataset of mice. This analysis successfully identified signature TEs of 2-cell-like cells (2CLCs) [21]. Employing MATES for quantifying TE expressions, we integrated the quantified TE count matrix with gene expression profiles, facilitating comprehensive clustering and visualization analyses, as shown in Fig.2 ab and Supplementary Fig.1 a. Our study revealed a distinct subpopulation of 2CLCs (cluster 17), positioned between stage II and stage III of reprogramming. Notably, MATES detected the 2CLCs

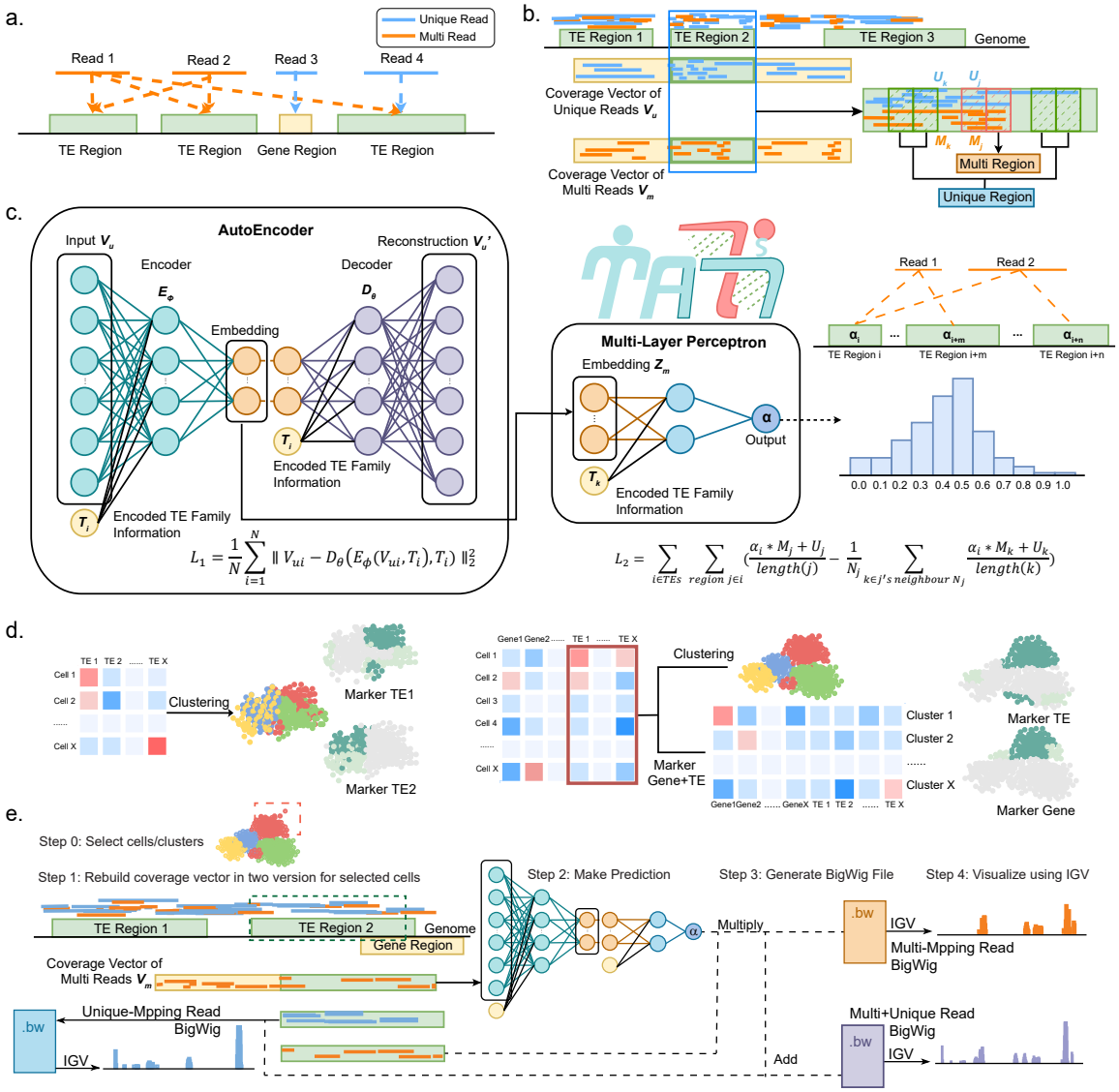

Fig. 1 MATES methodology for TE quantification and analysis. (a) Raw reads are aligned to the reference genome, accounting for multi-mapping reads at TEs' loci. (b) TE coverage vectors are constructed, capturing reads' distribution information. (c) An AutoEncoder model extracts latent embeddings from unique reads coverage vectors. These embeddings, combined with TE family data, predict the likelihood of multi-mapping reads aligning to each TE locus. (d) The multi-mapping probability α, computed by MATES, is critical in creating the TE count matrix. This matrix is pivotal for cell analyses and can be utilized either independently or in conjunction with a conventional gene count matrix. Such combined use enhances cell clustering and biomarker (gene and TE) discovery, providing a more comprehensive understanding of cellular characteristics. (e) Genome-wide reads coverage visualization by MATES in the Genome Browser. This method quantifies TEs at specific loci in individual cells, producing bigwig files with coverage from probabilistically assigned multi-mapping reads. These files, containing both unique and multi-mapping reads, are merged to generate comprehensive bigwig files for genome-wide TE read visualization using tools such as the Interactive Genomic Viewer (IGV).

population and distinguished their signature gene markers, especially Zscan4d and Zscan4c [22], within the transition-stage clusters. Moreover, MATES identified specific TE markers, MERVL-int and MT2_Mm, that are enriched within the 2CLC cluster, corroborating previous studies that recognize these TEs as defining markers for 2CLCs [23–26]. These findings highlight MATES's ability to capture cell populations and their significant biological markers (genes and TEs), providing insights into the cellular dynamics of reprogramming.

We next conducted a TE-centric analysis to validate further the distinct role of MATES's TE expression quantification in cell clustering and biomarker discovery (Fig.2 cd, Supplementary Fig.1 b). When quantifying TE expression, we took care to exclude overlapping regions between TEs and

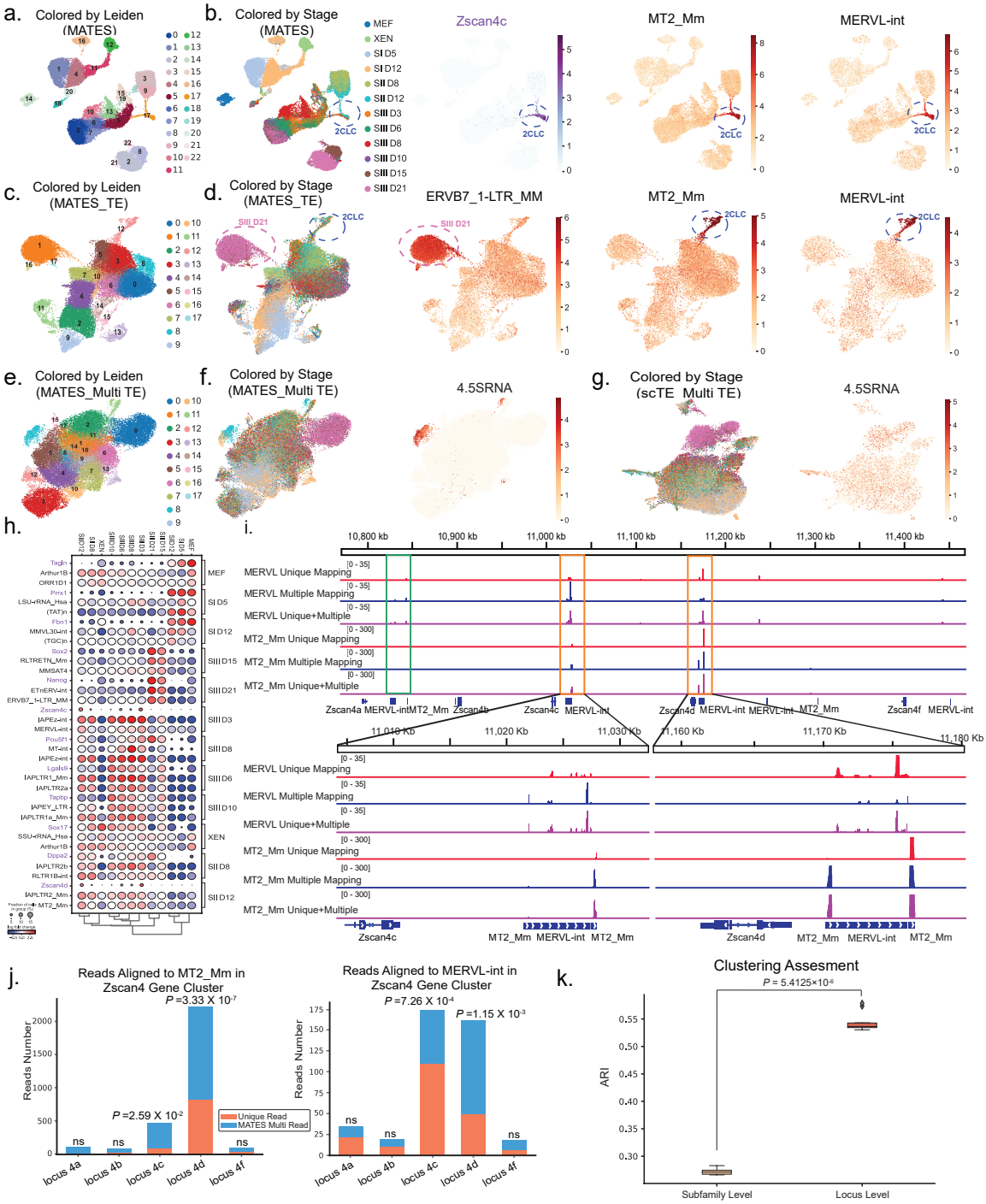

Fig. 2 MATES enhances cell clustering and biomarker discovery in mouse chemical reprogramming. (a-b) UMAP plots illustrate MATES's efficacy in cell clustering by integrating TEs and genes. (a) is colored by Leiden clustering results, while (b) is colored according to reprogramming stages, featuring identified gene (purple) and TE (red) biomarkers. (c-d) Additional UMAP plots emphasize MATES's capability for clustering using only TEs, with (c) colored by Leiden clusters, and (d) by reprogramming stages. Notably, MT2_Mm and MERVL-int TEs are prominent biomarkers in Zscan4c/Zscan4d-positive cells, consistent with known 2CLCs markers. (e-f) The UMAP plots demonstrate cell clustering derived from TE expression quantified from multi-mapping reads by MATES. Panel (e) displays cells colored according to Leiden clusters, while panel (f) shows the cells categorized and colored based on the various reprogramming stages. (g) UMAP plot of scTE quantified TE expression, colored by reprogramming stages, utilizing only multi-mapping reads. (h) Dot plot identifying stage-specific marker genes (purple) and TEs (black) as detected by MATES. (i) Illustration of MATES's probabilistic approach to allocating multi-mapping reads to specific TE loci, predominantly involving MT2_Mm and MERVL-int at the Zscan4c/Zscan4d loci (but not others) in 2CLCs (j) Bar plots displaying the reads enrichment for MT2_Mm and MERVL-int at the Zscan4c/Zscan4d loci. The enrichment p-value is calculated with a binomial test. (k) Box plot comparison of cell clustering efficacy between MATES's locus-level and subfamily-level TE quantification. The P-value was calculated using a one-sided Mann-Whitney U test.

their adjacent genes to prevent any potential information leakage from gene expression data. This TE-centric analysis specifically focused on TE expressions and it also identified the 2CLC cell population. Furthermore, this analysis not only confirmed the previous findings related to the 2CLCs population but also reaffirmed the relevance of its associated TE biomarkers, namely MERVL-int and MT2_Mm, as illustrated in Fig.2 cd. This shows that the robustness of our cell clustering and biomarker discovery is not merely an artifact of traditional gene expression analysis. Instead, TE quantification independently conducted by MATES proves capable of delivering solid cell clustering results and accurately capturing signature TEs for the identified cell populations. Moreover, by focusing on clustering driven by TE expression, quantified exclusively from multi-mapping reads to highlight MATES' proficiency in managing these reads, MATES adeptly pinpointed the previously mentioned biomarkers. It aligned them precisely with their specific developmental stages (Supplementary Fig.1 c). This approach also uncovered a distinct, small Cluster 8, marked by 4.5S RNA as its marker TE, a discovery missed by scTE (Fig.2 e-g).

Not only does MATES identify the signature genes and TE markers for 2CLCs and cell populations in different reprogramming stages (Fig.2 h, Supplementary Fig.1 d), but it also demonstrates a marked proficiency in accurately aligning the multi-mapping reads to specific loci, a challenge that has stymied current methods. For instance, scTE is limited to assigning multi-mapping reads to the metagene (TEs of the same subfamily), lacking a distinct assignment to specific genomic loci. While SoloTE quantifies reads uniquely mapped to TE at the locus level, for multi-mapping reads, it keeps only the best alignment and then quantifies at the subfamily level. In stark contrast, by leveraging the learned multi-mapping rate for each TE locus (α), MATES probabilistically assigns multi-mapping reads to TE genomic loci across the genome. By this strategy, we are able to quantify the TE expression at the locus level with accuracy, evident when assessing the multi-mapping reads for the 2CLC cells (Fig.2 i-k). Multi-mapping reads linked to MT2_Mm and MERVL-int were observed to align closely with genes Zscan4c and Zscan4d, and the total reads linked to MT2_Mm and MERVL-int that aligned closely with Zscan4c and Zscan4d loci are significantly higher than other control loci (Fig.2 ij, Supplementary Fig1.e). This alignment is consistent with findings by Zhu et al. [27], where the activation of Zscan4c was correlated with the activation of the endogenous retrovirus MT2/MERVL. Additionally, locus-specific TE quantification markedly improves clustering accuracy compared to the sub-family level TE quantification typically used in existing methods, clearly demonstrated in Fig.2 k. This emphasizes the substantial benefits of precise, locus-level TE quantification. For additional results demonstrating the effectiveness of MATES with this 10x single-cell RNA-seq data, please see Supplementary Fig.1.

MATES quantifies disease-related TE expression in a full-length Smart-Seq2 single-cell RNA-seq data of Human Glioblastoma.

To illustrate the comprehensive functionality of the MATES, we conducted an investigation using data from the Smart-Seq2 full-length single-cell RNA sequencing platform [28], particularly focusing on a human glioblastoma dataset [29]. This study involved quantifying TE expression and exploring its role in cellular heterogeneity in the Smart-Seq2 dataset. The combined use of MATES's TE expression quantification and conventional gene expression analysis allowed us to pinpoint distinct cell populations within the glioblastoma microenvironment, as shown in the UMAP plots (Fig.3 ab). Significantly, we observed certain TEs with expression patterns linked to crucial glioma gene markers like EGFR [29, 30] and TE markers including HUERS-P1-int [31], HERVK-int [32], along with immune cell gene markers such as CD74 [33, 34] and TE marker LTR2B [35, 36] (Fig.3 b). This correlation suggests TEs may play a significant role in the development of glioblastoma, potentially affecting both tumor heterogeneity and immune response. Combining TE-based cell typing with gene expression data revealed a detailed interplay between genes and TEs. This integration showcased how TE-based clustering could complement gene expression analysis, thus enhancing the resolution of cellular heterogeneity studies. Further demonstrating MATES's precision, we performed cell clustering solely based on the TE count matrix quantified by MATES. This approach successfully recaptured all major cell populations, pinpointing their associated TE markers (Fig.3 cd). The dot plots (Fig.3 e) not only displayed the associations between specific marker genes, TEs, and cell types but also quantified their relative expression levels, adding a deeper dimension to the data analysis.

In addition to analyzing TE expressions at the subfamily level, MATES' locus-level TE quantification offered a more comprehensive view of the cellular landscape (Fig.3 f-h, Supplementary Fig.2). This approach facilitated the identification of highly expressed TE loci corresponding to marker TEs

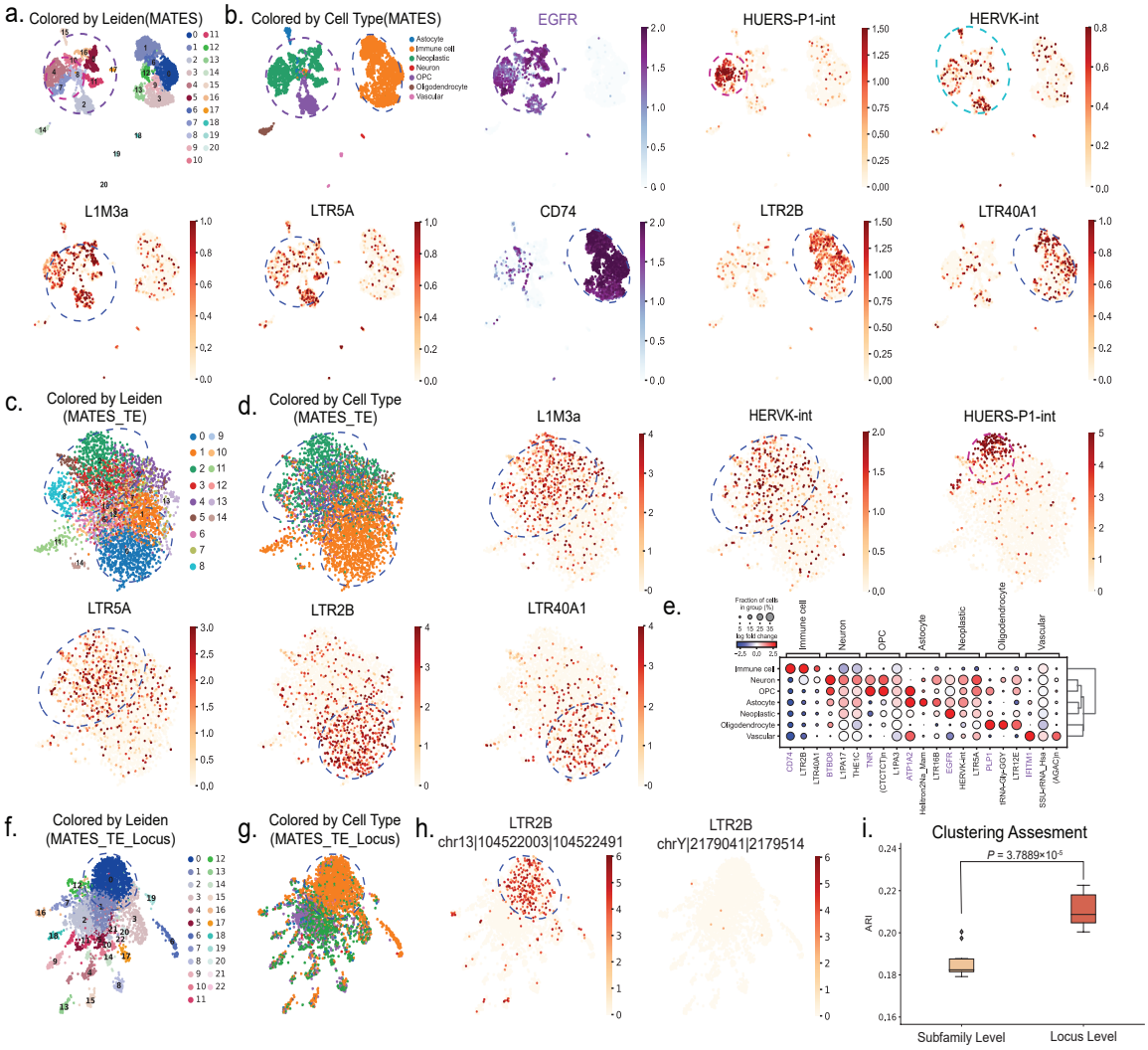

Fig. 3 MATES quantifies disease-related TE expression in a Smart-Seq2 single-cell RNA-seq data. (a-b) UMAP plots showcase cell clustering informed by gene and TE markers "MATES" or "Gene+TE" signifies combined gene expression with TE data quantified by MATES, "scTE" refers to scTE's quantification of gene and TE data, and "Gene" pertains to gene expression alone. Initially, MATES UMAPs are colored by Leiden clusters (a), then by cell type, glioma (EGFR, HUERS-P1-int, HERVHK-int, L1M3a, and LTR5A), and immune cell markers like CD74, LTR2B, and LTR40A1 (b). (c-d) UMAPs based solely on MATES-quantified TE expression are colored by Leiden clusters (c) and cell type with specific markers (L1M3a, HERVH-int, etc.) (d). (e) Dot plots elucidate the correlation between marker genes, TEs, and cell types, as identified by MATES. (f-h) Illustrate the enhanced clustering accuracy using MATES's locus-level TE quantification, with (f) detailing Leiden clusters and (g) showing cell types. (h) The plot lists a highly expressed TE marker (LTR2B) for Immune cells at the locus level and the same TE's nonopen locus, demonstrating MATES's nuanced capability in locus-level TE quantification. (i) A box plot comparison of cell clustering efficacy between MATES's locus-level and subfamily-level TE quantification reveals the method's enhanced resolution, demonstrating its performance in biomarker identification and cellular classification. P-value was calculated using a one-sided Mann-Whitney U test.

previously identified at the subfamily level. Notably, even for the same TE such as LTR2B, distinct loci could exhibit significantly varying expression levels (Fig.3 h), emphasizing the critical necessity for precise and locus-specific TE quantification. The application of this locus-specific TE quantification significantly improved cell clustering accuracy compared to its subfamily-level counterpart, as demonstrated in Fig.3 i. Thus, underscoring its substantial contribution to the analysis of cellular heterogeneity and the understanding of TE functions.

Our results affirm the robustness of MATES when applied to full-length scRNA-seq data, emphasizing its effectiveness for in-depth cellular analysis within these complex datasets. Notably, existing tools often struggle to handle this full-length single-cell RNA-seq data effectively, making MATES' capability to process and interpret such datasets even more valuable. This highlights the pivotal role of this approach in deepening our understanding of cellular heterogeneity in complex diseases such as

glioblastoma, especially in cases involving various sequencing platforms (e.g., full-length smart-seq2) where other tools like SoloTE encounter limitations.

MATES has broad applicability and can be utilized for TE quantification in single-cell data of other modalities

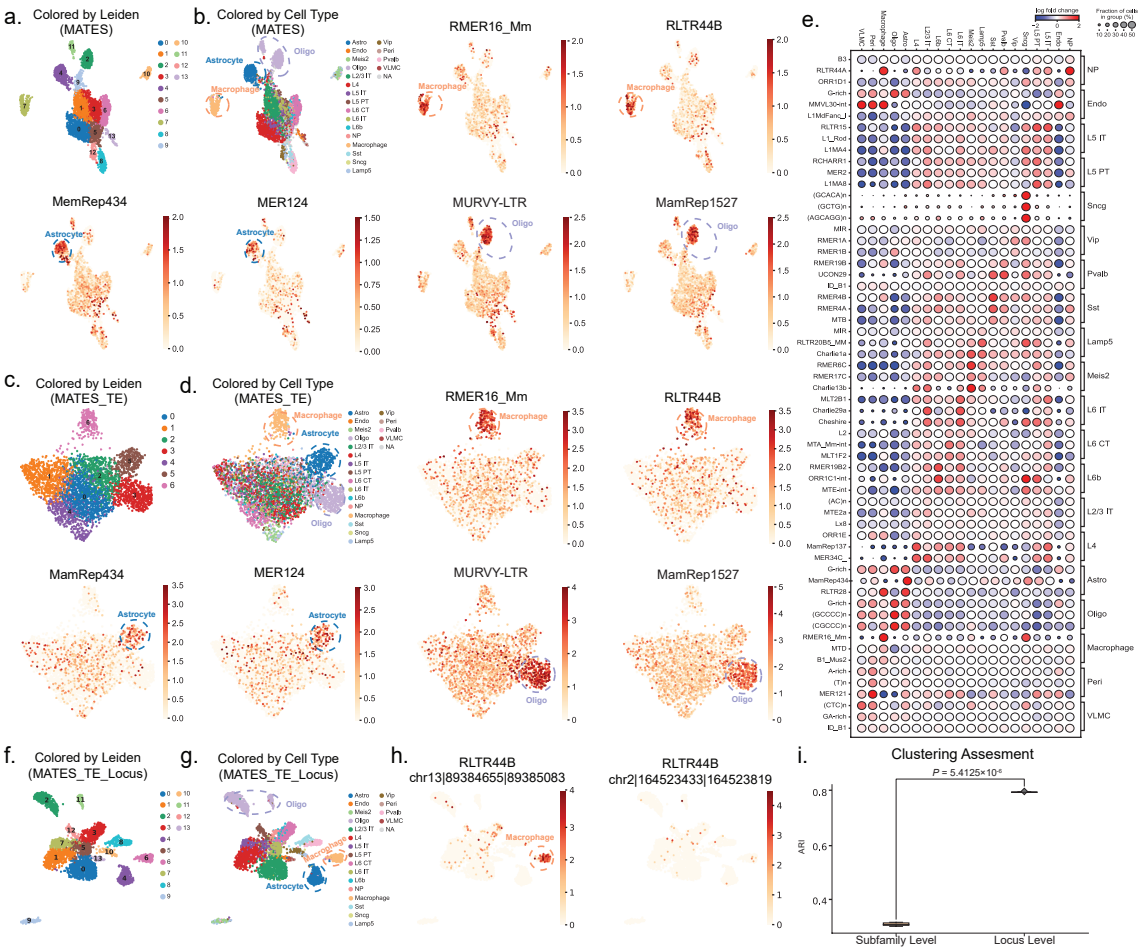

Fig. 4 MATES versatility application on adult mouse brain scATAC-seq data. (a-d) UMAP plots demonstrate the effectiveness of MATES quantification in cell clustering and identifying signature TE markers, using both TEs and Peaks for clustering, with (a) displaying Leiden clusters, and (b) illustrating cell types and TE markers. Key TE biomarkers such as RMER16_Mm, RLTR44B in macrophages, MemRep434, MER124 in Astrocytes, and MURVY-LTR, MamRep1527 in oligodendrocytes are identified. (c-d) Showcase the specificity of MATES in TE-centric clustering (using only TE quantification by MATES), with (c) focused on Leiden clusters, (d) on cell types and the distinctive TE markers previously noted. (e) Dot plot concisely presents cell type-specific TE markers uncovered by MATES. (f-h) These panels illustrate MATES's improved clustering accuracy using the locus level TE quantification. Panel (f) features UMAP visualizations based on locus-level TE quantification, with colors representing Leiden clusters. Panel (g) displays the same UMAP, but with color coding to differentiate various cell types. Panel (h) provides a specific example of the TE marker RLTR44B in Macrophages at the locus level, in contrast to a non-open locus of the same TE, demonstrating MATES's adeptness in detailed locus-level TE quantification. (i) The box plot in panel i contrasts the cell clustering efficacy of MATES when comparing locus-level versus subfamily-level TE quantification. This comparison highlights the advantages of employing locus-specific TE quantification. The associated P-value was determined using a one-sided Mann-Whitney U test.

Our proposed MATES model demonstrates its versatility by effectively quantifying TEs in not only transcriptome data but also epigenome data. To validate this adaptability, we applied MATES to a 10x single-cell ATAC-seq dataset of adult mouse brain [37]. This achievement is particularly

significant because many existing methods are tailored exclusively for transcriptomic data and may not perform as effectively with other single-cell modalities. MATES empowers the quantification of TE locus-specific attributes, such as chromatin accessibility, across diverse single-cell data modalities, extending its utility beyond transcriptomics.

By quantifying chromatin accessibility at the TE subfamily level and incorporating standard single-cell ATAC-seq peaks, MATES facilitates refined cell clustering and the identification of TE markers that are characteristic of distinct cell populations (Fig.4 ab, Supplementary Fig.3 a). Remarkably, TEs such as RMER16_MM and RLTR44B from the ERVK family show exclusive accessibility in macrophages [38, 39], while MamRep434 and MER124 are preferentially accessible in astrocytes [40], underscoring TEs' significant roles in neurogenesis and astrogliogenesis [41].

Leveraging solely the TE expression data quantified by MATES for clustering, we successfully identified TE biomarkers and maintained clear delineation among cell groups (Fig.4 cd, Supplementary Fig.3 b), confirming that the insights gleaned from TE quantification are rooted in genuine biological phenomena. This is reinforced by distinctive TE chromatin accessibility patterns, which provide a deeper understanding of the epigenomic landscape uniquely programmed in each cell type(Fig.4 e). For example, MamRep434's significant contribution to the motif of LHX2, a key transcription factor in neurogenesis, is emblematic of the functional implications of TE accessibility in cell identity and function [40, 42].

MATES' locus-specific TE quantification approach is instrumental in pinpointing signature TEs for each cell population with precision, providing exact locus positions. This fine-grained approach not only facilitates more accurate TE marker discovery but also enables rigorous experimental validation. The profound impact of this locus-specific TE expression quantification becomes apparent through its significant enhancement of cell clustering capabilities (Fig.4 f-i, Supplementary Fig.3 cd). For example, in the mouse brain scATAC dataset, the locus chr13|89384655|89385083|RLTR44B exhibits substantial chromatin accessibility in cells that should be annotated as microglia, despite their previous macrophage annotation (Fig.4 h). Within a 200Kb flanking region of this RLTR44B locus, the nearest gene is EDIL3. Quantitative proteomics have indicated increased EDIL3 expression in isolated microglia of APP-KI mice compared to WT mice [43]. Similarly, the locus chr19|40434224|40434373|MamRep434 (Supplementary Fig.3 c), located near the SORBS1 gene associated with astrogliosis, demonstrates elevated expression in individuals with schizophrenia who express high levels of inflammatory markers [44]. Furthermore, genes situated within a 200kb flanking region of this transposable element include ALDH18A1 [45] and ENTPD19 [46], which play pivotal roles in astrocyte functions and their interactions with non-astrocytic cells. These findings collectively underscore the regulatory potential of these identified marker TE loci in influencing astrocyte-related gene expression. Locus-level TE quantification demonstrates significantly higher cell clustering accuracy compared to subfamily-level analysis (Fig.4 i). This further underscores the effectiveness of MATES in locus-level TE quantifications and their potential benefits for understanding cellular states within the data. Marker TEs identified at the locus level align with those detected at the subfamily level, validating the method's accuracy and yielding locus-specific insights into TE's impact on chromatin accessibility [47]. Supplementary Fig.3 provides additional supporting evidence of MATES' effectiveness in the 10x scATAC-seq dataset, further solidifying its position as a versatile tool for TE quantification and analysis at the single-cell level across different modalities.

MATES empowers in multi-omics TE quantification and analysis

To further attest to the broad applicability of MATES, we applied it to a single-cell multi-omics dataset (10x multiome) [48]. By amalgamating the TE quantification provided by MATES with conventional gene expression data (RNA transcripts from scRNA-seq) and matched accessibility quantification (chromatin accessibility from scATAC-seq), we distinguished various cell populations and their associated markers, as depicted in Fig.5 ab.

MATES's ability to cluster diverse cells, identifying distinct cell populations and TE biomarkers across different sequencing techniques, introduced MATES' potential in harnessing multi-omics data for in-depth cell analysis (Fig.5 c-f), the results underscore the synergistic interplay of different modalities within MATES. Remarkably, when solely relying on MATES' TE quantification (TE_scRNA and TE_scATAC), the method still captured the primary cell populations and their signature TE markers. In contrast, quantification at the locus level further improved clustering performance(Fig.5 g-l, Supplementary Fig.4 a-d). Analyzing the multi-omics data, MATES unexpectedly reveals that certain TEs are uniquely discernible in specific cell sub-populations and exclusive to a particular modality

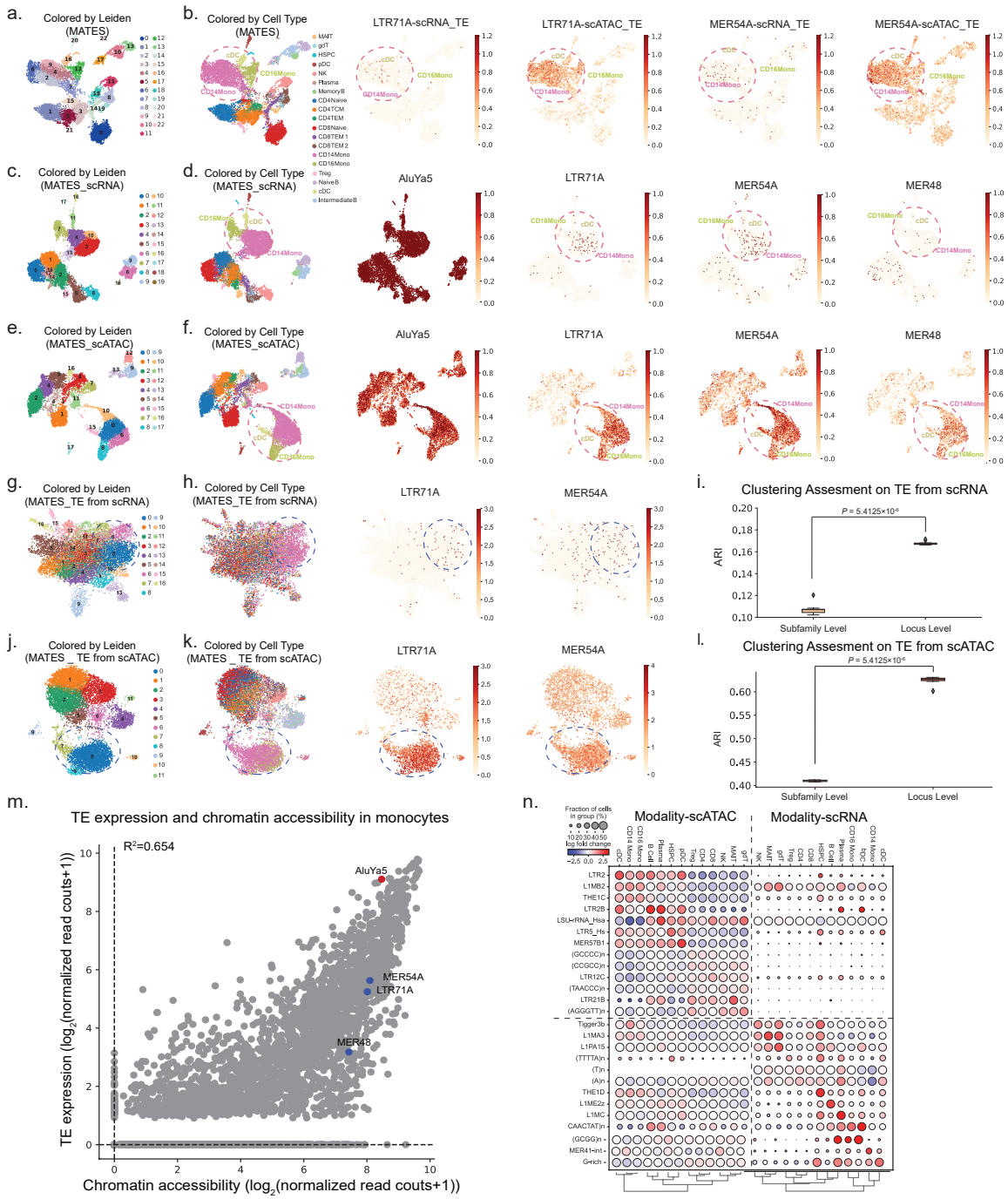

Fig. 5 Multi-modal TE analysis in human PBMCs using MATES. (a-b) UMAP plots of joint MATES clustering using both sc_RNA and sc_ATAC modalities, with (a) illustrating Leiden clusters and (b) detailing cell type clusters. (c-f) TE quantification across modalities highlights the complementarity of multi-modal TE quantification. Here, (c-d) feature UMAP clustering by Peaks and TE in the scATAC modality, while (e-f) display UMAP clustering by Genes and TE in the scRNA modality. Both (c) and (e) are colored by Leiden clusters, and (d) and (f) by cell types, showcasing the differential expression of TEs like AluYa5 across both modalities, whereas MER48, LTR71A, and MER54A appear specific to scATAC. (g-l) This series of UMAP plots and box plots illustrates the multi-modal TE analysis. (g) and (j) feature UMAP plots of TE expression, colored by Leiden clusters to highlight the clustering patterns. (h) and (k) are UMAP plots that focus on different cell types and TE markers, providing insights into the distinct identities of cells and their associated TEs. (i) and (l) are box plots contrasting the cell clustering effectiveness. The box plots underscore the improved resolution provided by locus-level quantification over subfamily-level quantification. P-values in these two plots were calculated using a one-sided Mann-Whitney U test. (m) Illustrates TE biomarkers discerned via scRNA and scATAC modalities, noting that high expression TEs often correlate with increased chromatin accessibility, while the reverse is not uniformly observed, highlighting the unique contributions of each modality. (n) A dot plot captures signature TEs per cell type, validating the complementary nature of scATAC and scRNA data for a holistic view of transposon dynamics.

(Fig.5 mn, Supplementary Fig.4 e). TEs with elevated gene expression often exhibit increased chromatin accessibility (represented by red dots, such as AluYa5, have been reported [49]). Conversely, transposons with enhanced chromatin accessibility do not always indicate high expression levels (marked as blue dots). For instance, LTR71A is predominantly detected via scATAC-seq, while not present in the scRNA-seq dataset. A similar trend is observed with TE biomarkers like MER48, and MER54A, suggesting these TEs are accessible in uninfected monocytes but transcriptionally dormant. This leads us to propose that such TEs may be primed to a "poised" state that might correlate to monocyte function [50]. Notably, several of these TEs, previously identified as "poised", featured prominently among the upregulated TEs. This enrichment calculated by hypergeometric test, with a significant p-value of $P = 5.91 \times 10^{-24}$, was determined by comparing the marker TEs identified through MATES with those upregulated TEs reported in the study (Supplementary Fig.4 f).

This observation highlights the capability of MATES to uncover novel biological features of TE. Here, our result emphasizes that when analyzed through the prism of TEs, chromatin accessibility, and RNA abundance offer complementary insights into the single-cell state (see Fig.5 m). The unique strength of MATES is demonstrated in Supplementary Fig.4 g. Within this framework, TE quantification by MATES boosts cell clustering accuracy and helps to build new hypotheses.

Method benchmarking in single-cell TE quantification

In our study, we conduct a detailed benchmarking of the MATES approach for Transposable Element (TE) quantification and its effect on cell clustering in single-cell datasets with various modalities. This is compared with the two established methods, scTE and SoloTE. Our benchmarking primarily concentrates on TE quantification at the sub-family level, due to the limitations of existing methodologies. Both scTE and SoloTE are incapable of providing locus-specific TE quantification that accounts for multi-mapping reads. Specifically, scTE is confined to sub-family level quantification, while SoloTE does offer locus-specific quantification but is constrained to unique mapping reads. The prevalence of multi-mapping reads among TEs highlights the critical need for accurate locus-specific TE quantification to enhance our understanding of cellular states. To ensure a fair comparison in our benchmarking, we have restricted our analysis in this section to the sub-family level, a feature shared by both scTE and SoloTE. However, the advantages of MATES in providing superior locus-specific TE quantification over the sub-family level have been discussed in earlier sections of our study for each individual dataset.

Here, we perform comprehensive benchmarking of TE quantification methods, evaluating them based on their downstream cell clustering performance. We analyze TE information from three distinct perspectives to illuminate the various aspects of TE expression's relevance and utility. Firstly, we analyzed gene/peak and TE expression concurrently to demonstrate how TE expression complements conventional gene/peak quantifications from single-cell data. This analysis underscores the added value that TE data can provide alongside traditional metrics. Secondly, we focused exclusively on TE expression to showcase the inherent power and information encapsulated within TE data alone. This approach aimed to illustrate that TE quantification, independent of conventional gene/-peak data, holds sufficient information for effective cell clustering. Lastly, we assessed TE expression based on multi-mapping reads to emphasize the superiority of the MATES method in handling these reads for TE quantification. This aspect was particularly crucial for understanding the robustness of MATES in dealing with the complexities of TE data. These comparisons were referred to as Gene/Peak + TE expression (Gene+TE/Peak+TE), TE expression (TE), and Multi-mapping TE expression (Multi TE), respectively.

The benchmarking in this work included single-cell RNA datasets obtained using both the 10x and Smart-Seq2 platforms, as well as single-cell ATAC sequencing (scATAC) datasets. It is important to note that SoloTE is not compatible with Smart-Seq2 and scATAC single-cell data. Consequently, SoloTE was excluded from the comparative analysis for these two datasets. To assess the clustering results derived from the three perspectives outlined earlier, we employed the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) as our key benchmarking standards. Our analysis compared MATES with two existing single-cell TE quantification methods scTE and SoloTE, which primarily differ in their approach to handling multi-mapping reads. We observed significant differences in their performance: scTE uses a basic correction method, SoloTE selects the top-scoring alignment, and MATES utilizes a probabilistic approach informed by local read context. The statistical significance of our results, indicated by p-values, reinforces the superiority of MATES in TE quantification, as evidenced by improved cell clustering performance compared to existing methods.

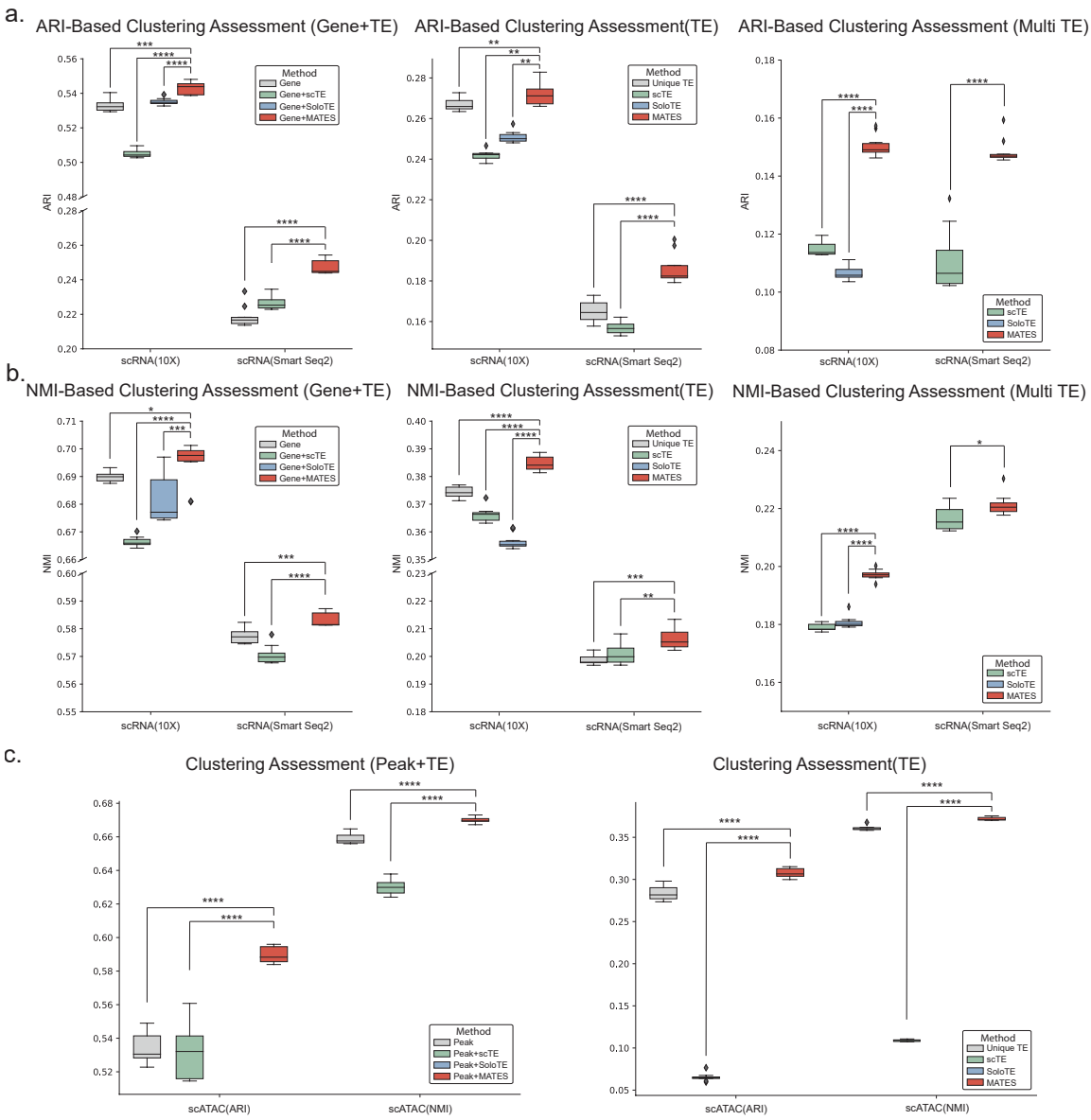

Fig. 6 Benchmarking MATES performance in diverse single-cell datasets This assessment highlights MATES's efficiency in cell clustering, evaluated through the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), using different TE quantification strategies. (a-b) The impact of various TE quantification methods on cell clustering is compared within the 10x scRNA Dataset of Chemical Reprogramming and Smart-Seq2 Dataset of Glioblastoma. These methods include scTE (gene expression combined with scTE-quantified TE), SoloTE (gene expression combined with SoloTE-quantified TE), and MATES (gene expression integrated with MATES-quantified TE). MATES outperforms both scTE and SoloTE by enhancing gene expression with TE data (left). The middle panel compares clustering based solely on TE quantification methods—unique TEs, scTE, SoloTE, and MATES—with 'unique TEs' representing unique-mapping reads TE expression, highlighting MATES's consistent superiority. The right panel confirms MATES's advantage over scTE and SoloTE when considering only multi-mapping TE reads. Note: SoloTE's incompatibility with Smart-Seq2 data results in a blank section. Panel (a) uses ARI for evaluation, while panel (b) utilizes NMI. (c) The 10x scATAC Dataset of the Adult Mouse Brain is analyzed to contrast peak and TE quantification using scTE and MATES against peak-only datasets (left). TE mapping reads from scTE and MATES are also compared against unique TE mapping reads (right). SoloTE's incompatibility with scATAC data leads to its exclusion from this part of the analysis. (The P-values were calculated using a one-sided Mann-Whitney U test, with significance levels indicated as * for $P < 0.05$, ** for $P < 0.01$, *** for $P < 0.001$, and **** for $P < 0.0001$).

In the analysis that integrated both Gene and TE expression, MATES demonstrated superior performance over scTE and SoloTE, especially evident in the 10x scRNA dataset for chemical reprogramming. In this context, scTE's effectiveness was markedly inferior compared to a gene expression-only approach, highlighting the importance of accurate multi-mapping read assignment. This was evidenced by MATES achieving higher ARI and NMI scores across all datasets tested, as shown in the left panels of Fig.6 ab. To further showcase and compare the accuracy of TE quantification by different methods, we conducted a cell clustering analysis based on only the TE quantification, excluding gene expression data. Here again, MATES outshone scTE and SoloTE in terms of clustering efficiency, which was reflected in improved ARI and NMI scores (see the middle panels of Fig.6 ab). This result underscores the capability of MATES in handling multi-mapping reads and demonstrates its significance in TE quantification. In scenarios focusing exclusively on multi-mapping TE reads, the proficiency of MATES in managing these challenging read assignments became apparent. MATES consistently outperformed scTE and SoloTE in these tests, as depicted in the middle and right panels of Fig.6 ab. This consistent superiority across different testing conditions demonstrates the robustness and effectiveness of MATES in TE quantification. Its particular strength in the assignment of multi-mapping reads highlights its proficiency, significantly contributing to the accuracy in subsequent cell clustering tasks.

In our benchmarking analysis of the 10x scATAC mouse brain dataset, we assessed the MATES approach against scTE and peak-only quantification methods. This evaluation aimed to determine the cell clustering accuracy when combining peak and TE data. Our results indicated a clear advantage of MATES over both scTE and peak-only methods in terms of clustering accuracy, showcasing its capacity to integrate additional chromatin accessibility insights effectively. A key aspect of MATES's performance was its comprehensive quantification of TE expression, which encompassed both unique mapping reads and overall TE data. This approach significantly surpassed scTE, contributing to an enhanced clustering process. The inclusion of multi-mapping reads by MATES, although less critical than in scRNA datasets due to its lower frequency in this scATAC-seq data, provided valuable insights to the clustering analysis, as illustrated in Fig.6 c. These findings not only highlight the effectiveness of MATES but also demonstrate its adaptability and potential for broad applications in various single-cell data modalities.

Validation of MATES quantification accuracy with single-cell long-read sequencing data

To ascertain the precision of TE quantification by MATES, we utilized a melanoma brain metastasis dataset, as detailed in Shiau et al. [51]. This dataset comprised single-cell nanopore RNA sequencing (scNanoRNAseq) [52], producing long-read data with an average length of 937 base pairs (bps). This is significantly longer than the corresponding short-read data of 222 bps obtained from 10x scRNA-seq (real sequencing reads), which came from the same set of cells in the study. The longer read length of nanopore sequencing is advantageous for TE quantification, as it allows for more accurate read alignment[53] (i.e., reduced multi-mapping rate $\tilde{1}\%$ in nanopore sequencing compared to over 12% in 10x short-read sequencing from the same study), thereby enhancing the reliability of TE quantification. This level of precision from long-read data serves as the ground truth for our validation of short-read TE quantification by MATES.

We applied the methods on the matched short-read data from the same set of cells using 10x sequencing to quantify the TE expression (at both sub-family and locus levels). This provided a direct means to validate the quantification accuracy of MATES on short-read data against the ground truth established by the long-read nanopore sequencing. Furthermore, to rigorously evaluate MATES's performance with even shorter reads, we simulated 100 bps reads in full-length format from the corresponding nanopore long reads. For these simulated short reads, we already knew the ground-truth mapping locus, i.e., the locus mapped to by the corresponding long read. These shorter reads introduce a heightened challenge due to an increased potential for multi-mapping issues. The methodology and reasoning behind this simulation are detailed in the Methods section. By using both long and significantly shorter reads from the same cell set, we conducted an extensive evaluation of the effectiveness of MATES in TE quantification in comparison to other existing methods.

In our validation of MATES through nanopore long-read sequencing data, we compared TE expression quantified by each method within individual cells, calculating their correlation using R^2. This analysis is illustrated in panels a and b of Fig.7. To enhance the accuracy of our correlation calculations and accurately reflect the relationship between short-read and long-read expressions,

outliers (denoted as grey dots) were excluded, as detailed in the methods section. This preprocessing step was crucial in ensuring that the R^2 values truly represented the biological data, free from distortions caused by non-expressive or anomalous readings. In this validation, MATES's quantification of TEs demonstrated a strong correlation with the ground-truth long-read data ($R^2 = 0.7218$), surpassing both scTE ($R^2 = 0.6384$) and SoloTE ($R^2 = 0.2509$) in quantifying TE expression at the sub-family level from the real 10x short-read data, as seen in Fig.7 a. Interestingly, the correlation of TE expression quantified by SoloTE significantly decreased (from 0.7078 to 0.2509) when including multi-mapping reads (refer to Supplementary Fig.5 a), underscoring the importance of accurate multi-mapping read assignment. Additionally, the pseudobulk TE expression in the data, determined by averaging the TE expression across all cells in the real short-read 10x sequencing dataset, is also quantified by each respective method. These pseudobulk TE expression were then correlated with the ground-truth pseudobulk TE expression derived from the nanopore long-read single-cell sequencing data. This correlation was quantified in terms of the coefficient of determination (R^2), as shown in Supplementary Fig.5 b. In this figure, each data point represents a pseudobulk TE expression value, calculated using all cells across the entire dataset. In a comparative analysis, MATES demonstrated a stronger correlation with the established ground-truth long-read data, achieving an R^2 value of 0.9248. This performance exceeded that of other methods such as scTE and SoloTE, which achieved R^2 values of 0.8988 and 0.7789, respectively.

Further experiments with simulated short-read single-cell data confirmed the robustness of MATES. Given SoloTE's incompatibility with the full-length single-cell RNA-seq data, we evaluated the correlation of TE expression quantified by MATES and scTE in the simulated short-read data against nanopore long-read data quantification. MATES consistently outperformed scTE, achieving an R^2 of 0.8331 compared to scTE's R^2 of 0.1500, as shown in Fig.7 b. The pseudobulk TE expression from the simulated short read dataset is illustrated in Supplementary Fig.5 c, demonstrating that MATES outperforms scTE, with respective R^2 values of 0.9187 and 0.1964.Further examination of scTE's results revealed anomalously high expression in specific TEs, such as SSU-rRNA_Hsa and LSU-rRNA_Hsa. Excluding these TE loci led to more accurate results($R^2 = 0.8628$), potentially indicating that bias toward certain TE families largely contributes to scTE's lower overall performance, detailed in Supplementary Fig.5 c.

In the final stage of our validation process, we sought to assess MATES's proficiency in quantifying locus-specific TE expression using simulated short-read single-cell datasets. For the actual 10x short-read single-cell data, we had corresponding long-read data from nanopore single-cell sequencing. However, a direct mapping of each short-read in the 10x data to its specific long-read in the nanopore dataset was not feasible, which made it challenging to accurately validate the exact genomic loci for the read assignments in the 10x short-read dataset and thus the locus-level TE quantification. Nevertheless, locus-specific validation is feasible with simulated short-read datasets, where each simulated read is paired with a corresponding long read from the nanopore single-cell data. Our analysis of the simulated short-read dataset compares MATES with the best alignment score strategy for multimapping reads assignment, depicted in Fig.7c, MATES showed a robust correlation (with $R^2 = 0.6093$) with the reference standard of locus-specific TE quantification obtained from nanopore long-read sequencing compare to best alignment score strategy (with $R^2 = 0.1548$). This result underscores MATES's proficiency in precisely profiling locus-specific TE expression. It highlights MATES's capability to deliver accurate and comprehensive TE quantification at both the individual locus and sub-family levels. This is especially vital for conducting detailed single-cell genomics research, focusing particularly on the role and impact of transposons.

Discussion

To address the multi-mapping read assignment challenge in accurate TE quantification, here we introduce MATES, a deep neural network framework tailored for quantifying TE expressions at the single-cell level, with a particular focus on multi-mapping TE reads. Utilizing an AutoEncoder, MATES learns the distribution patterns of unique mapping reads at individual TE loci based on the coverage vectors from unique-dominant TE regions. It integrates both unique-mapping and multi-mapping reads to accurately quantify TE expression at the locus level. When applied to various datasets, including 10x and Smart-Seq2 scRNA-seq data, 10x scATAC-seq data, and 10x Multiome data, it consistently outperformed existing methods. This tool is not limited to subfamily-level TE quantification but also delivers accurate locus-level quantification, thus enhancing the resolution of analysis for cell populations and leading to the identification of novel locus-specific TE markers.

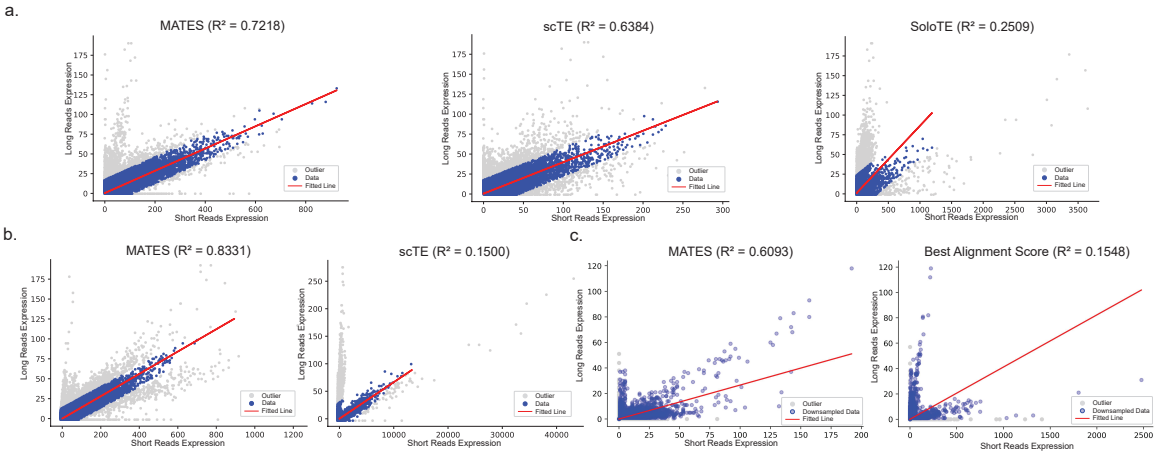

Fig. 7 Validation of MATES TE quantification using nanopore long-read single-cell sequencing. (a) Analyze the correlation between TE expression quantified by MATES, scTE, and SoloTE from real 10x short-read data and the nanopore long-read sequencing for the same set of cells. This comparison is at the sub-family level. (b) Examines the correlation of sub-family level TE expression quantified by MATES and scTE from simulated short-read data with the ground truth TE quantification from the nanopore long-read sequencing data. (c) Evaluate the correlation of locus-level TE expression quantified by MATES on the simulated short-read data and the ground truth locus-level TE quantification provided by the nanopore long-read sequencing.

MATES stands out in the realm of single-cell TE quantification and analysis, offering distinct characteristics and functionalities, notably its proficient handling of multi-mapping TE reads. Central to MATES is a deep learning model adept at analyzing the distribution of background reads mapped to TE loci and their surrounding areas. This in-depth understanding of background read distribution is crucial for accurately assigning reads to TE regions, thereby enhancing the clarity of genetic regulatory activities within biological systems. Distinctively, MATES achieves locus-level TE quantification for individual cells, a step beyond existing methods that typically provide sub-family level TE quantification or omit multi-mapping reads altogether. This approach leverages the inherent heterogeneity of single-cell data. By doing so, MATES enables comprehensive exploration of locus-specific interactions between genes and TEs across varied cell populations, yielding valuable insights into their unique contributions to broader biological processes. Moreover, MATES' compatibility with a variety of single-cell sequencing platforms and diverse omics data types significantly broadens its applicability and supports multi-omics analysis of the TE system. This versatility fosters the potential for novel biological insights, enabling researchers to glean detailed information from multi-omics data. In addition, MATES brings enhanced functionality with its whole-genome TE quantification and visualization capabilities within specific cell populations. The ability to visualize whole-genome TE expressions in selected cell groups, utilizing tools like IGV, facilitates the rapid identification of potential gene-TE interactions. Complemented by an online visualization interface, MATES offers an accessible and efficient platform to analyze and interpret complex TE-related data. This interface simplifies the exploration of TE expressions and interactions, allowing researchers to intuitively navigate and interpret the genomics landscape. By integrating interactive features and user-friendly tools, MATES eases the process of identifying gene-TE interactions and their implications across different cellular contexts, thereby enriching the field of genomics and epigenetics.

As described above, the methodological innovations of MATES has led to interesting biological discoveries for further experimental examinations. In the context of chemically reprogramming mouse cells to an early embryonic-like state, the MATES framework has shown remarkable proficiency in identifying TE markers. Significantly, markers such as MT2 and MERVL-int were enriched in the Zscan4 gene cluster, especially flanking Zscan4c and Zscan4d genes in 2CLC cells. This analysis also uncovered potential TE markers at various stages of reprogramming, underscoring the model's ability to detect TE markers amidst the dynamic cellular reprogramming processes. When applied to human glioblastoma data, MATES successfully distinguished TE markers specific to cancer cells and immune cells, demonstrating its potential in identifying target TEs, which could contribute to advancements in cancer treatment. Beyond scRNA data, MATES was also applied to adult mouse brain data to consider chromatin accessibility in TE regions. This application revealed that the ERVK TE family

is highly expressed in cells like macrophages and astrocytes, highlighting the versatility and broad applicability of the MATES framework in diverse biological contexts. Surprisingly, employing MATES to sc-multi-omics data uncovers the TEs' epigenetic poised state, which underscores the advantage MATES supplied here for multi-modalities analysis.

While MATES, a deep neural network-based method, typically requires more computational resources such as GPUs than its counterparts scTE and SoloTE for quantifying Transposable Elements (TEs), its memory usage remains relatively stable (generally below 30,000M) across single-cell datasets of varying sizes. This is attributed to its batch-based model training. Additionally, MATES's running time is not significantly impacted by the number of cells in a dataset. For datasets of approximately 2000 to 5000 cells, MATES can complete TE quantification within 200 minutes on a computer with specific hardware configurations (Intel Xeon Platinum 8160 CPU @ 2.10GHz, NVIDIA A16), excluding time for read preprocessing and mapping (Supplementary Fig. 6 a). This sublinear increase in time and memory requirements is partly due to our downsampling strategy used in model training, as not all TE instances from every cell are necessary for training. This approach allows MATES to efficiently process large-scale single-cell datasets without being constrained by memory limitations or prohibitive running times. However, the total runtime, including read preprocessing and mapping, tends to increase linearly with data size. Overall, the total running time of MATES is generally comparable to that of scTE and SoloTE (Supplementary Fig. 6b).

In summary, the introduction of MATES likely stands to have a profound impact on the field of single-cell genomics. It brings into focus the intricate dynamics of TEs, converting what was once considered 'junk reads' into a treasure trove of information that can redefine our understanding of cellular diversity and dynamics. As single-cell technologies evolve, accurate and precise TE quantification will become increasingly crucial. MATES is poised to be an indispensable asset in this endeavor, guiding researchers through the genomic complexities to discover new facets of biology at the single-cell level.

Methods

Raw reads preprocessing

Raw reads (fastq files) are mapped to the genome using STAR [54] for Smart-Seq2 scRNA-seq and STAR-solo [55] for 10x scRNA-seq, and the mapped reads are segregated into unique mapping reads that align to a single locus, and multiple mapping reads that potentially map to multiple genome locations (Fig.1 a).

Following alignment, samtools [56] is used to separately extract unique mapping reads and multi-mapping reads from the alignment files (i.e., bam) and store them into separate bam files. The unique mapping reads are identified by checking if MAPQ equals 255, and multi-mapping reads are extracted according to the bam file attribute NH, which indicates the number of loci the reads map to. For 10x genomics sequencing data, we have an extra step to build individual bam files for each cell according to barcodes from bam files. We then use these bam files to construct two distinct read coverage vectors (V_u and V_m) centered on the individual TE regions. V_u represents the coverage vector of unique mapping reads, while V_m denotes the coverage vector of multi-mapping reads around the TE regions. This procedure is built based on pysam and pybedtools [56–59]. In this context, "coverage" refers to the count of sequencing reads mapped to a specific genomic location. These coverage vectors serve to model the expression pattern of TEs.

To comprehensively cover Transposable Element (TE) regions, our method involves constructing a coverage vector centered at the start position of each TE region (locus). Given that over 97% of TE regions are typically under 1000 base pairs (bp) in length, as illustrated in Supplementary Fig.7 ab, we set a maximum TE region length (S) of 1000 bp. Accordingly, we define the coverage vector to span 2001 bp, extending 1000 bp both upstream and downstream from the TE start position. This strategy ensures that not only is the entirety of the TE region captured in most instances but also a minimum of 1000 bp of flanking genomic sequence is included for context. For the minority of TE loci exceeding 1000 bp (approximately 3%), we truncate these regions to the initial 1000 bp. This truncation is crucial to maintain uniformity in the input vectors for the neural network, ensuring that each input vector corresponds to the same number of neurons in the network's input layer.

In our approach, we address the challenge of overlapping regions between Transposable Elements (TEs) and nearby genes. To tackle the issue of double-counting reads and the risk of information leakage at the intersections of TE and gene regions, we incorporate a specific preprocessing step.

This step, executed using the bedtools[58], involves trimming the TE reference against the gene reference alignment. Such trimming effectively removes overlaps between TEs and genes, reducing the likelihood of inflated TE read counts due to these intersections. However, it's important to recognize that this preprocessing may not always be preferable in every application. While we have applied it in our current study to ensure that the contribution of TEs to downstream analyses is accurate and not confounded by overlapping genes, in some practical scenarios, maintaining the natural structure of these regions might be more beneficial. Trimming overlaps with gene regions could unintentionally distort TE quantification, particularly in cases where TEs intersect with neighboring genes. Therefore, in real-world applications, opting out of this preprocessing step could be advantageous. By doing so, the natural intersections of genes and TEs are preserved, allowing for a more authentic representation of their influence on TE quantification. Moreover, for those who wish to use the unaltered references, this option remains available in MATES, providing flexibility to accommodate different research needs and perspectives.

Unique-dominant and Multi-dominant Bins

The key of our method is the accurate attribution of multi-mapping reads to specific TE loci, achieved by estimating the multi-mapping reads ratio α_i for each TE locus i. This estimation relies on the distribution of reads (i.e., read coverage vectors) around each TE locus. In a given cellular context, we choose TE regions that encompass both uniquely mapped and multi-mapped reads from the same cell as our training samples. Subsequently, we divide the TE locus (region) into small bins of a length W (a hyperparameter set to 5 base pairs by default). Given that read distribution across a region of any gene or TE locus tends to be smooth, we postulate that neighboring regions in close proximity within the same TE locus should have similar read distributions. This assumption is well-supported by the fact the coverage difference between neighboring bins within the same TE locus is close to zero as we observed in our single-cell data (Supplementary Fig.7 c). Please be aware that this histogram is derived from TE loci predominantly composed of unique reads with confident mapping (TEs characterized solely by U bins), thereby avoiding significant multi-mapping uncertainty and ensuring reliable read coverage. We then employ this supported assumption to provide constraints for multi-mapping read assignments, thus incorporating it as a part of the loss term for training our deep neural network.

To ascertain the proportions of multi-dominant (M) and unique-dominant (U) bins within each TE region, we apply a W bp slicing window. M and U bins are determined by the percentage of unique and multi-mapping reads within each bin. A bin is categorized as an M-bin if the percentage of multi-mapping reads exceeds a threshold P; otherwise, it is deemed a U-bin. Selection of the hyperparameters, the bin size (W) and proportion threshold (P), is dependent on the data's multi-mapping rates. For data with high multi-mapping rates, we opt for a smaller bin size and a higher threshold to enhance accuracy. Conversely, for low multi-mapping rate data, we prefer a larger bin size and a lower threshold to acquire more training samples (please see Supplementary Fig.7 de for the selection of those parameters). The motivation to segregate bins into U and M categories stems from the desire to utilize the more reliable U bins (where unique reads are assigned confidently) to constrain the estimation of the multi-mapping rate within a specific TE locus (Fig.1, b).

Estimating the multi-mapping rate for each TE locus via a deep neural network

TE loci are divided into three specific groups based on the presence of unique-dominant (U) and multi-dominant (M) bins. The groups comprise TE regions that contain both M and U bins; TE regions composed solely of M bins; and TE regions that consist only of U bins. Our model utilizes TE regions from the first group—those containing both M and U bins—as training samples. The presence of both types of bins in these regions is necessary for calculating the loss used in optimizing our neural network. It's important to highlight, however, that certain TE families may have a higher degree of repetitiveness, leading to a higher occurrence of M bins in their instances. Consequently, our training dataset, curated from single-cell data, might lean towards an overrepresentation of these specific TE families. To counteract this potential imbalance, we have implemented a downsampling strategy, where we limited training samples for each TE family not exceeding 2500 inputs (Larger values are recommended to enhance TE quantification performance, although this will result in extended training time). This approach aims to decrease the overrepresentation of certain TE families,

thereby fostering a more balanced representation within our model. We also removed TE families with too little training sample (less than 50 samples) to focus on more representative TE families. Such balanced representation is instrumental in enhancing the model's overall performance and accuracy. Moreover, our downsampling strategy results in the size of the training sample being more dependent on the number of expressed TE families in the dataset than on the total number of cells. This method helps in maintaining both training time and memory usage at a steady level, regardless of the cell count.

Next, we employ coverage vectors constructed from unique mapping reads V_u to train the AutoEncoder(AE) model [60]. Recognizing that different TE families may exhibit distinct expression patterns, we introduce TE family information into the input and embedding layers to account for such heterogeneity. This is accomplished by encoding the TE family information of each TE region using one-hot encoding. The AE model minimizes the Mean Squared Error loss function (L_1) to minimize the discrepancy between the input vector and the reconstructed vector, resulting in a more meaningful reduced embedding of the TE locus (Fig.1 c).

$$L_1 = \frac{1}{N}\sum_{i=1}^{N} \|V_{ui} - D_\theta(E_\phi(V_{ui}, T_i), T_i)\|_2^2 \tag{1}$$

Here N represents the number of training samples, V_{ui} represents the unique mapping reads coverage vector of training sample TE locus i, while the T_i is its one-hot encoded TE family information for the specific TE locus. E_θ and D_ϕ represents encoding function and decoding function respectively. Subsequently, the trained AE model generates embeddings to represent the read distribution of the TE loci in the latent space. These embeddings are combined with the TE family information of the corresponding TE loci and utilized as the input for a Multilayer perceptron (MLP) regressor [61] (Fig.1 c). For the MLP, we define the loss function (L_2) to minimize the read coverage difference between each M-bin (j) and its nearest neighboring U-bins (N_j). The loss function (L_2) is defined as follows:

$$L_2 = \sum_{i \in TEs} \sum_{j \in i} \left(\frac{\alpha_i * M_j + U_j}{L_j} - \frac{1}{N_j} \sum_{k \in N_j} \frac{\alpha_i * M_k + U_k}{L_k} \right) \tag{2}$$

Here, M_j and U_j represent the total number of multi-mapping reads and unique mapping reads falling within the j-th M-bin of TE locus i, while L_j represents the number of covered base pairs in this region. Similarly, M_k and U_k represent the total depth of the k-th U bin among the N_j nearest neighboring U bins of the j-th M bin, and L_k represents the number of covered base pairs in that $k-th$ bin. The output α_i represents the probability that a multi-mapping read falls into this TE locus. The grand loss to train the entire model is the sum of both loss terms $L_{grand} = L_1 + \lambda * L_2$, where λ is a hyperparameter.

TE quantification and characterization at the single-cell level

In our study, we employ a neural network framework that integrates an Autoencoder (AE) with a Multi-layer Perceptron (MLP) regressor. This hybrid model is adept at assigning multi-mapping reads to specific TE loci, a crucial step in quantifying TE expression in individual cells. The process begins by feeding the model a read coverage vector. This vector, which is derived from unique-mapping reads, is specific to a given TE locus i, denoted as V_{ui}, accompanying this vector is the TE family information associated with the TE instance. Our model, leveraging its neural network capabilities, computes a multi-mapping probability, α_i. This probability reflects the likelihood of multi-mapping reads being associated with the TE locus i. Subsequently, this computed probability is utilized to estimate the count of multi-mapping reads for the TE locus in question. It achieves this by multiplying α_i, with the aggregate of all potential multi-mapping reads within that particular TE locus i. The outcome of this multiplication is an estimated count of multi-mapping reads assigned to the specific TE locus. The total read number of a given cell at TE locus i is then calculated by $UR_i + \lceil \alpha_i \times MR_i \rceil$, where UR_i is the unique read number and the MR_i is the multi-read number at the TE locus i respectively. To facilitate a comprehensive understanding of TE expression at the cellular level, we construct a cell-by-TE expression matrix. This matrix is a confluence of two critical components: the estimated counts of multi-mapping read (as determined by our neural network model) and the counts of uniquely mapped read corresponding to the same TE locus. We have meticulously designed this matrix to integrate seamlessly with the conventional cell-by-gene expression matrices. This compatibility is

pivotal for enabling effective downstream analyses. One such downstream analysis is cell clustering, which is significantly enhanced by incorporating our cell-by-TE expression matrix. By using this matrix for cell clustering, we can underscore the effectiveness and precision of our TE quantification method. An illustrative example of this process and its outcomes is presented in Fig.1 d.

After identifying distinct cell populations in the single-cell data with TE quantification from MATES, we further employ the biomarker calling method from SCANPY [62], which systematically identifies signature genes and Transposable Elements (TEs) that show higher expression in specific cell populations compared to other cells. Its primary function is to distinguish signature genes and TEs unique to each cell population.

MATES enhances its functionality by producing TE reads alignment files in BigWig format. This process involves a series of steps. Initially, coverage vectors are recalculated for each TE locus, denoted as i. This recalculation is done separately for unique reads and multi-mapping reads, using a comprehensive TE reference that includes all TEs, without any trimming or exclusion of TE loci overlapping with genes. Subsequently, at each base pair within a TE locus, the coverage from multi-mapping reads is adjusted. This adjustment involves multiplying the read coverage of the multi-mapping reads at each specific base pair by the locus-specific multi-mapping ratio, α_i. The final read coverage at each base pair is then determined by summing the coverage from unique reads and the adjusted coverage from multi-mapping reads. The generated bigwig files include all reads assigned probabilistically to TE regions by our neural network and are compatible with genomic visualization tools such as IGV [63] (see Fig.1 e), facilitating genome-wide TE examinations and visualizations. To provide a comprehensive exploration of the outputs from MATES, including cell clusters, gene biomarkers, TE biomarkers, and the probabilistic assignment of reads across the genome, we have developed an interactive MATES web server, available at https://mates.cellcycle.org. This platform allows users to interact with and analyze the data generated by our method, offering a resource for detailed investigation of TEs and genes across different cell populations.

MATES application to single-cell data across different modalities

In order to demonstrate the broad applicability of MATES, we evaluated this approach on single-cell datasets across diverse modalities and multiple platforms. These include 10x scRNA-seq, Smart-Seq2 scRNA-seq, 10x scATAC-seq, and 10x multiome. We implemented a consistent processing sequence for these datasets, regardless of their platform or modality.

For single-cell RNA-seq datasets derived from platforms like 10x and Smart-Seq2, we adhered to the aforementioned procedure. When conducting single-cell ATAC-seq analysis, we chose STAR-solo for read mapping due to its ability to handle multi-mapping reads, unlike traditional single-cell ATAC-seq read mapping tools such as CellRanger ATAC which discard these multi-mapping reads [64]. We then applied MATES to the bam files generated by STAR-solo, leading to the creation of a cell-by-TE chromatin accessibility matrix. Concurrently, we generated the conventional cell-by-peak count matrix using our chosen tool, in this case, the CellRanger ATAC toolset. Nonetheless, users have the flexibility to perform conventional peak quantification using any tool of their preference. This matrix was integrated with the cell-by-TE matrix derived from MATES via concatenation, enabling subsequent downstream analyses like cell clustering. In cases where users engage with other single-cell modalities not discussed in this study, such as single-cell DNA methylation profiling, they can utilize MATES to generate the cell-by-TE matrix for downstream analysis. However, the extraction of cell-by-gene or cell-by-peak matrices connected to the cells can be achieved through established pipelines, which fall outside the purview of our TE-focused study.

For the processing of multi-omics datasets, we employed a multi-step strategy. Initially, we performed independent alignments for each modality, using the STAR-Solo alignment tool and adhering to the specific considerations previously mentioned. Using the MATES method, We subsequently quantified individual TE for each modality. Following this, we consolidated the TE expressions from all modalities with their respective gene/peak expressions, quantified through conventional methods such as Cell Ranger. We then concatenated the cell-by-TE matrices across modalities or integrated them using other multi-modal data integration methods, like Seurat[65]. The resulting integrated multi-omics TE matrices offered by MATES provide a comprehensive representation of the cells, leading to enhanced cell clustering and identification accuracy. Furthermore, multi-omics gene and TE quantification allow us to explore the roles of TEs in the identified cell populations from different modality perspectives, thus yielding deeper biological insights.

Method benchmarking design

To benchmark the accuracy of our Transposable Element (TE) quantification method, we focused on evaluating its impact on cell clustering performance within single-cell datasets. The rationale behind this approach is based on the hypothesis that more accurate TE quantification leads to enhanced cell clustering and identification. This enhancement in cell clustering can be quantitatively assessed using two key metrics: the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Both metrics are crucial in single-cell research, as they provide a means to compare clustering results against known cell type labels, offering a clear indication of clustering accuracy. Our benchmarking process involved comparing the performance of our method with other state-of-the-art TE quantification methods, scTE and SoloTE. SoloTE is known for its ability to quantify unique read TE expression at the locus level and multi-mapping read TE expression at the subfamily level, while scTE quantifies both at the subfamily level. To ensure a consistent and fair comparison, we normalized the data by aggregating locus-level TE expressions to subfamily levels across all methods. This standardization is vital for making meaningful comparisons between the methodologies. The effectiveness of each method in improving cell clustering performance was statistically evaluated using the one-sided Mann–Whitney U test. This test determines whether the improvements in ARI and NMI scores with a particular TE quantification method are statistically significant compared to others.

In our study, we sought to showcase the added value of Transposable Elements (TEs) as an informative component in addition to traditional gene expression analysis. To achieve this, we conducted cell clustering analyses under different scenarios: using only gene expression, combining gene and TE expressions, and using TE expression alone. The clustering based on gene expression alone established a baseline for comparing the enhanced performance achieved when TE data was incorporated using the three methods. Additionally, to specifically understand the impact of TE expression on clustering performance, we performed clustering analyses solely based on TE expression. This allowed us to isolate the effect of TE data from the gene expression influence. To create a foundational comparison point, we utilized a unique-mapping reads-only approach for TE quantification. This approach, which omits multi-mapping reads, provided a clearer perspective on the efficacy of each method's TE quantification process. Furthermore, we evaluated the effectiveness of TE expression quantification when relying exclusively on multi-mapping reads. This was crucial to demonstrate the capability of our method, particularly in handling the complexities of multi-mapping scenarios through context-aware expression analysis. MATES underwent extensive testing across various single-cell modalities and platforms, highlighting its versatility and broad applicability in different research contexts.

Validation with nanopore long-read sequencing data

To validate the accuracy of our TE quantification model, we employed nanopore long-read sequencing data, specifically scNanopore-seq [52], recognized for its minimal multimapping issues due to the length of long reads and hence serving as a reliable ground truth for TE expression measurement. We applied both our method and other benchmarked methods to the 10x short-read sequencing data from the same study, corresponding to the same cell sets analyzed via the scNanopore-seq approach. The focus was on comparing the TE expression levels quantified by MATES, scTE, and SoloTE from the short-read data against the long-read data, considered as the 'ground truth'. For this comparison, we used the coefficient of determination (R^2) as the metric to assess the accuracy of TE quantification by each method. A higher R^2 value, indicating greater agreement with the long-read ground truth, would suggest a more precise TE quantification from the matched short-read data by each method. This approach allowed us to evaluate the accuracy of TE quantification methods using the nanopore long-read sequencing data as a benchmark, providing a robust validation of our model against other existing techniques. We further broadened the scope of our validation process by simulating single-cell sequencing data with even shorter reads than those used in the 10x single-cell data. We generated short reads of 100 base pairs (bp) from the long-read data using SeqKit [66]. This approach aimed to mimic scenarios with shorter reads, which typically result in a higher likelihood of multi-mapping issues. However, due to SoloTE's incompatibility with full-length single-cell RNA-seq data, our performance comparison on these simulated 100 bp short reads was confined to scTE and MATES

In our methodology for calculating the coefficient of determination (R^2), the initial step was to remove TEs that were not expressed in any cells from both long-read and short-read datasets. After this initial filtering, we applied linear regression to the data that remained. Our analysis, conducted at the TE subfamily level, included a filtering phase where we computed the standard deviation of the

residuals. These residuals represent the differences between the predicted and actual TE expression values. To enhance the robustness of our analysis, data points where residuals exceeded five standard deviations from the mean were excluded, effectively removing significant outliers. This step was crucial to improve the precision and reliability of our R^2 calculation by minimizing the impact of extreme outliers. In Fig.7, this refined dataset is visually represented with data points in blue, outliers in grey, and the regression line in red. Furthermore, we calculated the median TE expression across all cells, known as pseudobulk TE quantification, applying the same exclusion criteria. We then calculated R^2 for this pseudobulk TE quantification to evaluate its accuracy against the ground truth pseudobulk TE quantification obtained from nanopore long-read sequencing. This comparison was done using various methods for short-read data quantification, as detailed in Supplementary Fig.5 bc. This comprehensive approach allowed for a detailed evaluation of both individual cell-level and aggregate TE quantification accuracy at the subfamily level.

For the real 10x short-read single-cell sequencing data, the absence of corresponding long-reads from matched nanopore single-cell sequencing precludes locus-specific validation. To assess accuracy at the locus level, we analyzed the coefficient of determination (R^2) using our simulated short-read single-cell dataset, where short reads were derived from the long reads of the nanopore single-cell data, providing a direct correspondence necessary for calculating locus-level TE quantification correlation accurately. In line with standard practices for handling multimapping reads, such as those used in SoloTE, our analysis included a traditional strategy that focuses on reads with the highest alignment scores in multi-mapping scenarios. Following the data exclusion criteria similar to those used in the TE sub-family level correlation analysis, the remaining data points were employed to compute R^2 for both MATES and the traditional method. Given the substantial volume of data points, which complicates visualization, each pseudo point in Fig.7 c represents three actual data points. The color density in this figure reflects the concentration of data points, facilitating visual interpretation of data distribution and density. However, the actual calculation of R^2 was conducted using all data points, ensuring a comprehensive assessment of locus-level quantification accuracy.

Data availability

The TE reference data was sourced from the Repeat Masker website. Specifically, the Mouse TE reference is available at https://www.repeatmasker.org/species/mm.html, and the Human TE reference can be accessed at https://www.repeatmasker.org/species/hg.html. In this study, we utilized a range of publicly accessible benchmark datasets that cover scRNA-seq, scATAC-seq, and multi-omics datasets across various biological contexts. These datasets include the scRNA-seq 10x mouse chemical reprogramming dataset (GSE114952) as detailed in Zhao et al. 2018 [21], the Smart-Seq2 human glioblastoma dataset (GSE84465) described in Darmanis et al. 2017 [29], the scATAC-seq 10x 5k adult mouse brain cell dataset [37] available at https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k, and the 10x Multi-Omics 10k PBMC dataset [48] accessible at https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. Additionally, the study utilized full-length mRNAs sequenced by single-cell nanopore (GSE212945) as detailed in Shiau et al. 2023 [51].

Software availability

MATES, developed as a Python-based tool, is publicly accessible and comes with comprehensive documentation. It can be found at https://github.com/mcgilldinglab/MATES. Additionally, for enhanced user interaction, especially in visualizing TE quantifications and cell clustering, a user-friendly web interface has been established. This can be accessed at https://mates.cellcycle.org.

Acknowledgments

This work is supported by grants from the Canadian Institutes of Health Research (CIHR) [PJT-180505 to J.D]; the Funds de recherche du Québec - Santé (FRQS) [295298 to J.D., 295299 to J.D.]; the Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN2022-04399 to J.D.]; and the Meakins-Christie Chair in Respiratory Research [to J.D.]; and Tao P. Wu is a CPRIT scholar for cancer research, supported by CPRIT award (RR180072). We thank MyCellome LLC for supporting the visualization of MATES.

References

[1] Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., *et al.*: Ten things you should know about transposable elements. Genome biology **19**, 1–12 (2018)

[2] He, J., Babarinde, I.A., Sun, L., Xu, S., Chen, R., Shi, J., Wei, Y., Li, Y., Ma, G., Zhuang, Q., *et al.*: Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scte. Nature communications **12**(1), 1456 (2021)

[3] Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., *et al.*: A single-cell atlas of in vivo mammalian chromatin accessibility. Cell **174**(5), 1309–1324 (2018)

[4] Lanciano, S., Cristofari, G.: Measuring and interpreting transposable element expression. Nature Reviews Genetics **21**(12), 721–736 (2020)

[5] Berrens, R.V., *et al.*: Locus-specific expression of transposable elements in single cells with cello-seq. Nature Biotechnology **40**, 546–554 (2022)

[6] Garcia-Perez, J.L., Widmann, T.J., Adams, I.R.: The impact of transposable elements on mammalian development. Development **143**(22), 4101–4114 (2016)

[7] Li, C., Zhang, Y., Leng, L., Pan, X., Zhao, D., Li, X., Huang, J., Bolund, L., Lin, G., Luo, Y., *et al.*: The single-cell expression profile of transposable elements and transcription factors in human early biparental and uniparental embryonic development. Frontiers in Cell and Developmental Biology **10**, 1020490 (2022)

[8] Bonté, P.-E., Arribas, Y.A., Merlotti, A., Carrascal, M., Zhang, J.V., Zueva, E., Binder, Z.A., Alanio, C., Goudot, C., Amigorena, S.: Single-cell rna-seq-based proteogenomics identifies glioblastoma-specific transposable elements encoding hla-i-presented peptides. Cell Reports **39**(10) (2022)

[9] Yang, W.R., Ardeljan, D., Pacyna, C.N., Payer, L.M., Burns, K.H.: Squire reveals locus-specific regulation of interspersed repeat expression. Nucleic acids research **47**(5), 27–27 (2019)

[10] Bendall, M.L., De Mulder, M., Iñiguez, L.P., Lecanda-Sánchez, A., Pérez-Losada, M., Ostrowski, M.A., Jones, R.B., Mulder, L.C., Reyes-Terán, G., Crandall, K.A., *et al.*: Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. PLoS computational biology **15**(9), 1006453 (2019)

[11] Rodríguez-Quiroz, R., Valdebenito-Maturana, B.: Solote for improved analysis of transposable elements in single-cell rna-seq data using locus-specific expression. Communications Biology **5**(1), 1063 (2022)

[12] Treangen, T.J., Salzberg, S.L.: Repetitive dna and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics **13**(1), 36–46 (2012)

[13] Goerner-Potvin, P., Bourque, G.: Computational tools to unmask transposable elements. Nature Reviews Genetics **19**(11), 688–704 (2018)

[14] Schwarz, R., Koch, P., Wilbrandt, J., Hoffmann, S.: Locus-specific expression analysis of transposable elements. Briefings in Bioinformatics **23**(1), 417 (2022)

[15] Shema, E., Bernstein, B.E., Buenrostro, J.D.: Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nature Genetics **51**(1), 19–25 (2019)

[16] Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., Greenleaf, W.J.: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**(7561), 486–490 (2015)

[17] Lee, J., Hyeon, D.Y., Hwang, D.: Single-cell multiomics: technologies and data analysis methods. Experimental & Molecular Medicine **52**(9), 1428–1442 (2020)

[18] Hu, Y., An, Q., Sheu, K., Trejo, B., Fan, S., Guo, Y.: Single cell multi-omics technology: methodology and application. Frontiers in cell and developmental biology **6**, 28 (2018)

[19] Baysoy, A., Bai, Z., Satija, R., Fan, R.: The technological landscape and applications of single-cell multi-omics. Nature Reviews Molecular Cell Biology, 1–19 (2023)

[20] Vandereyken, K., Sifrim, A., Thienpont, B., Voet, T.: Methods and applications for single-cell and spatial multi-omics. Nature Reviews Genetics, 1–22 (2023)

[21] Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., Chen, S., Jiao, Z., Xu, X., Xu, J., *et al.*: Single-cell rna-seq reveals dynamic early embryonic-like programs during chemical reprogramming. Cell stem cell **23**(1), 31–45 (2018)

[22] Falco, G., Lee, S.-L., Stanghellini, I., Bassey, U.C., Hamatani, T., Ko, M.S.: Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Developmental biology **307**(2), 539–550 (2007)

[23] Eckersley-Maslin, M.A., Svensson, V., Krueger, C., Stubbs, T.M., Giehr, P., Krueger, F., Miragaia, R.J., Kyriakopoulos, C., Berrens, R.V., Milagre, I., *et al.*: Mervl/zscan4 network activation results in transient genome-wide dna demethylation of mescs. Cell reports **17**(1), 179–192 (2016)

[24] Fu, X., Djekidel, M.N., Zhang, Y.: A transcriptional roadmap for 2c-like–to–pluripotent state transition. Science advances **6**(22), 5181 (2020)

[25] Yang, M., Yu, H., Yu, X., Liang, S., Hu, Y., Luo, Y., Izsvák, Z., Sun, C., Wang, J.: Chemical-induced chromatin remodeling reprograms mouse escs to totipotent-like stem cells. Cell Stem Cell **29**(3), 400–418 (2022)

[26] Sakashita, A., Kitano, T., Ishizu, H., Guo, Y., Masuda, H., Ariura, M., Murano, K., Siomi, H.: Transcription of mervl retrotransposons is required for preimplantation embryo development. Nature genetics **55**(3), 484–495 (2023)

[27] Zhang, W., Chen, F., Chen, R., Xie, D., Yang, J., Zhao, X., Guo, R., Zhang, Y., Shen, Y., Göke, J., *et al.*: Zscan4c activates endogenous retrovirus mervl and cleavage embryo genes. Nucleic acids research **47**(16), 8485–8501 (2019)

[28] Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R.: Full-length rna-seq from single cells using smart-seq2. Nature protocols **9**(1), 171–181 (2014)

[29] Darmanis, S., Sloan, S.A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., *et al.*: Single-cell rna-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell reports **21**(5), 1399–1410 (2017)

[30] Voena, C., Di Giacomo, F., Panizza, E., D'amico, L., Boccalatte, F., Pellegrino, E., Todaro, M., Recupero, D., Tabbo, F., Ambrogio, C., *et al.*: The egfr family members sustain the neoplastic phenotype of alk+ lung adenocarcinoma via egr1. Oncogenesis **2**(4), 43–43 (2013)

[31] Schneider, J., Volkmer, I., Engel, K., Emmer, A., Staege, M.S.: Expression of a new endogenous retrovirus-associated transcript in hodgkin lymphoma cells. International Journal of Molecular Sciences **20**(21), 5320 (2019)

[32] Shah, A.H., Rivas, S.R., Doucet-O'Hare, T.T., Govindarajan, V., DeMarino, C., Wang, T., Ampie, L., Zhang, Y., Banasavadi-Siddegowda, Y.K., Walbridge, S., et al.: Human endogenous retrovirus k contributes to a stem cell niche in glioblastoma. The Journal of clinical investigation **133**(13) (2023)

[33] Su, H., Na, N., Zhang, X., Zhao, Y.: The biological function and significance of cd74 in immune

diseases. Inflammation Research **66**, 209–216 (2017)

[34] Xu, S., Li, X., Tang, L., Liu, Z., Yang, K., Cheng, Q.: Cd74 correlated with malignancies and immune microenvironment in gliomas. Frontiers in Molecular Biosciences **8**, 706949 (2021)

[35] Deniz, Ö., Ahmed, M., Todd, C.D., Rio-Machin, A., Dawson, M.A., Branco, M.R.: Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. Nature communications **11**(1), 3506 (2020)

[36] Katoh, H., Honda, T.: Roles of human endogenous retroviruses and endogenous virus-like elements in cancer development and innate immunity. Biomolecules **13**(12), 1706 (2023)

[37] Genomics, X.: Fresh Cortex from Adult Mouse Brain (P50). https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k Accessed 2019-11-21

[38] Saito, T., Miyagawa, K., Chen, S.-Y., Tamosiuniene, R., Wang, L., Sharpe, O., Samayoa, E., Harada, D., Moonen, J.-R.A., Cao, A., *et al.*: Upregulation of human endogenous retrovirus-k is linked to immunity and inflammation in pulmonary arterial hypertension. Circulation **136**(20), 1920–1935 (2017)

[39] Gruchot, J., Herrero, F., Weber-Stadlbauer, U., Meyer, U., Kuery, P.: Interplay between activation of endogenous retroviruses and inflammation as common pathogenic mechanism in neurological and psychiatric disorders. Brain, Behavior, and Immunity **107**, 242–252 (2023)

[40] Sekine, K., Onoguchi, M., Hamada, M.: Transposons contribute to the acquisition of cell type-specific cis-elements in the brain. Communications Biology **6**(1), 631 (2023)

[41] Evering, T.H., Marston, J.L., Gan, L., Nixon, D.F.: Transposable elements and alzheimer's disease pathogenesis. Trends in Neurosciences **46**(3), 170–172 (2023)

[42] Lattke, M., Goldstone, R., Ellis, J.K., Boeing, S., Jurado-Arjona, J., Marichal, N., MacRae, J.I., Berninger, B., Guillemot, F.: Extensive transcriptional and chromatin changes underlie astrocyte maturation in vivo and in culture. Nature communications **12**(1), 4335 (2021)

[43] Xue, F., Du, H.: Trem2 mediates microglial anti-inflammatory activations in alzheimer's disease: lessons learned from transcriptomics. Cells **10**(2), 321 (2021)

[44] Catts, V.S., Wong, J., Fillman, S.G., Fung, S.J., Shannon Weickert, C.: Increased expression of astrocyte markers in schizophrenia: association with neuroinflammation. Australian & New Zealand Journal of Psychiatry **48**(8), 722–734 (2014)

[45] Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.*: Model-based analysis of chip-seq (macs). Genome biology **9**(9), 1–9 (2008)

[46] Ulivieri, C., De Tommaso, D., Finetti, F., Ortensi, B., Pelicci, G., D'Elios, M.M., Ballerini, C., Baldari, C.T.: At cell suppressive circuitry mediated by cd39 and regulated by shcc/rai is induced in astrocytes by encephalitogenic t cells. Frontiers in Immunology **10**, 1041 (2019)

[47] Zu, S., Li, Y.E., Wang, K., *et al.*: Single-cell analysis of chromatin accessibility in the adult mouse brain. Nature **624**, 378–389 (2023)

[48] Genomics, X.: PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting (10k). https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k Accessed 2020-09-09

[49] Groza, C., Chen, X., Pacis, A., Simon, M.-M., Pramatarova, A., Aracena, K.A., Pastinen, T., Barreiro, L.B., Bourque, G.: Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. Cell Genomics **3**(5) (2023)

[50] Chen, X., Pacis, A., Aracena, K.A., Gona, S., Kwan, T., Groza, C., Lin, Y.L., Sindeaux, R.,

Yotova, V., Pramatarova, A., et al.: Transposable elements are associated with the variable response to influenza infection. Cell Genomics **3**(5) (2023)

[51] Shiau, C.-K., Lu, L., Kieser, R., Fukumura, K., Pan, T., Lin, H.-Y., Yang, J., Tong, E.L., Lee, G., Yan, Y., *et al.*: High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. Nature communications **14**(1), 4124 (2023)

[52] Wang, Y., Zhao, Y., Bollas, A., Wang, Y., Au, K.F.: Nanopore sequencing technology, bioinformatics and applications. Nature biotechnology **39**(11), 1348–1365 (2021)

[53] Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q.: Opportunities and challenges in long-read sequencing data analysis. Genome biology **21**(1), 1–16 (2020)

[54] Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**(1), 15–21 (2013)

[55] Kaminow, B., Yunusov, D., Dobin, A.: Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data. Biorxiv, 2021–05 (2021)

[56] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The sequence alignment/map format and samtools. bioinformatics **25**(16), 2078–2079 (2009)

[57] Heger, A., Belgrad, T., Goodson, M., Jacobs, K.: pysam: Python interface for the SAM/BAM sequence alignment and mapping format. https://github.com/pysam-developers/pysam (2023)

[58] Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)

[59] Dale, R.K., Pedersen, B.S., Quinlan, A.R.: Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics **27**(24), 3423–3424 (2011)

[60] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006)

[61] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks **2**(5), 359–366 (1989)

[62] Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)

[63] Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**(2), 178–192 (2013)

[64] Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.*: Massively parallel digital transcriptional profiling of single cells. Nature communications **8**(1), 14049 (2017)

[65] Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., Satija, R.: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology (2023) https://doi.org/10.1038/s41587-023-01767-y

[66] Shen, W., Le, S., Li, Y., Hu, F.: Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PloS one **11**(10), 0163962 (2016)