Nucleic Acids Research, 2024, **52**, e93
https://doi.org/10.1093/nar/gkae793
Advance access publication date: 13 September 2024
Methods

IRescue: uncertainty-aware quantification of transposable elements expression at single cell level

Benedetto Polimeni[1,2], Federica Marasca[1], Valeria Ranzani[1,*] and Beatrice Bodega[1,2,*]

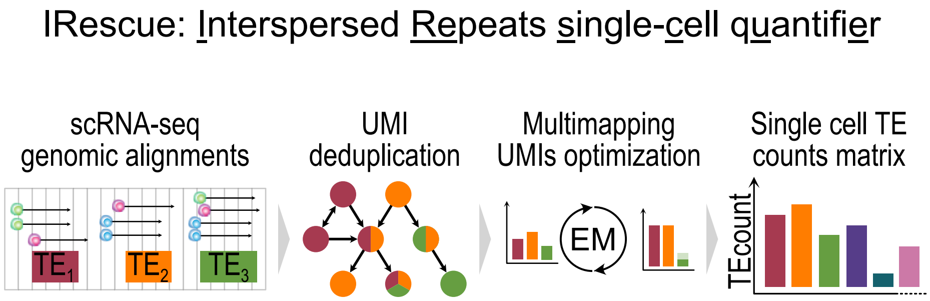
[1]INGM, Istituto Nazionale di Genetica Molecolare 'Romeo ed Enrica Invernizzi', Milan, Italy
[2]Department of Biosciences, University of Milan, Milan, Italy

[*]To whom correspondence should be addressed. Tel: +39 02 00 660 302; Fax: +39 02 00 660 216; Email: bodega@ingm.org
Correspondence may also be addressed to Valeria Ranzani. Tel: +39 02 00 660 345; Fax: +39 02 00 660 216; Email: ranzani@ingm.org

Abstract

Transposable elements (TEs) are mobile DNA repeats known to shape the evolution of eukaryotic genomes. In complex organisms, they exhibit tissue-specific transcription. However, understanding their role in cellular diversity across most tissues remains a challenge, when employing single-cell RNA sequencing (scRNA-seq), due to their widespread presence and genetic similarity. To address this, we present IRescue (Interspersed Repeats single-cell quantifier), a software capable of estimating the expression of TE subfamilies at the single-cell level. IRescue incorporates a unique UMI deduplication algorithm to rectify sequencing errors and employs an Expectation-Maximization procedure to effectively redistribute the counts of multi-mapping reads. Our study showcases the precision of IRescue through analysis of both simulated and real single cell and nuclei RNA-seq data from human colorectal cancer, brain, skin aging, and PBMCs during SARS-CoV-2 infection and recovery. By linking the expression patterns of TE signatures to specific conditions and biological contexts, we unveil insights into their potential roles in cellular heterogeneity and disease progression.

Graphical abstract


IRescue: Interspersed Repeats single-cell quantifier
scRNA-seq genomic alignments
UMI deduplication
Multimapping UMIs optimization
Single cell TE counts matrix
TE1
TE2
TE3
EM
TEcount

Introduction

Transposable elements (TEs) are mobile genetic elements present in the genome of most eukaryotes, constituting around 46% of the human genome (1). TEs can be hierarchically categorized into classes (such as LINE, SINE, LTR), families (like LINE1, Alu, ERVL) and subfamilies (for instance, L1PA2, AluY, HERVL) (2,3). Besides the existence of full-length elements that permit TE mobilization, the genome is invaded by TE fossils inserted in proximity to or into introns of the majority of genes, resulting from retrotransposition events that had occurred throughout the evolution (4,5). Nevertheless, they can still be transcribed within adjacent transcriptional units, providing regulatory elements that influence gene expression and RNA processing (6,7). TEs exhibit tissue-specific transcription patterns (8) and transcripts originating from TEs play a role in the epigenetic regulation of cell identity and differentiation (9,10). Next Generation Sequencing (NGS) technologies were indispensable to identify and annotate TEs in reference genomes, shedding light on the impact of TEs within genomic and transcriptional regulatory networks. However, the repetitive nature and substantial homology between elements present challenges in the NGS-based study of TEs (11). Implementing certain precautions in library design, such as opting for a paired-end layout, extending read length (12) and utilizing specific software (13,14), can significantly enhance both the mappability of reads and the accuracy of expression estimates for TEs. Although numerous tools are available for bulk RNA-Seq analysis (15–18), there have been limited efforts in developing methods for quantifying TE expression in single-cell RNA sequencing (scRNA-seq) datasets (19–21). The majority of scRNA-seq libraries in public repositories originate from droplet-based technologies (e.g. Chromium

Received: December 20, 2022. **Revised:** August 22, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** September 2, 2024

10x, Drop-seq and inDrops kits) (22), and are consequently characterized by short reads with a pronounced 3′- or 5′-end positional bias. Additionally, these reads are effectively single-end, as only one mate represents the cDNA insert, while the other carries the cell barcode and unique molecule identifier (UMI) sequences. The use of tag-based library layouts reduces read mappability and complicates the precise determination of the genomic origin of RNA fragments containing transposable elements. Existing tools for quantifying TE subfamilies in UMI-based scRNA-seq data typically employ a single alignment per UMI, often selected randomly, neglecting information from ambiguous alignments across different TE subfamilies (19,20). Moreover, these tools do not address sequencing errors in UMI sequences, which are common and can lead to overestimating UMI counts (23). In recent studies, researchers have started exploring the expression of TEs within the framework of single-cell heterogeneity in multicellular organisms (19,24,25). Nonetheless, the existing state-of-the-art methods lack the sensitivity required for accurately quantifying multi-mapping scRNA-seq reads. This limitation hinders the comprehensive discovery of distinct TE expression patterns. Here we present IRescue (Interspersed Repeats single-cell quantifier), a command-line tool designed for the deduplication and quantification of UMIs mapped onto TEs in scRNA-seq. In comparison to other freely available tools for quantifying TE subfamilies in scRNA-seq (19,20), IRescue incorporates gold-standard procedures for UMI deduplication that consider UMI frequencies and account for sequencing errors. It employs a probabilistic assignment method using an Expectation-Maximization (EM) algorithm to redistribute counts from multi-mapping reads to distinct TE subfamilies. We demonstrate the precision of IRescue by evaluating its performance on both simulated and real single-cell and single-nuclei RNA sequencing data. We analyze TE expression in colorectal cancer, human brain, aging and SARS-CoV-2 infection and recovery, revealing insights into their heterogeneity. Our results indicate that IRescue stands out as the most accurate tool for quantifying TE expression at the subfamily level.

Materials and methods

IRescue workflow

Input and output data

The only input file required to run IRescue is a binary aligned map (BAM) file (26) containing read alignments on a reference genome, with cell barcode and UMI sequence annotated as user-defined BAM tags ('CB' and 'UR' by default). A BAM file with these requirements can be obtained by widely used spliced aligners for scRNA-seq, such as Cell Ranger (27) or STARsolo (28,29). The genomic coordinates of TEs can be retrieved automatically by IRescue from the Repeatmasker annotation hosted at the UCSC servers, by simply indicating the name of the genome assembly on the command line (e.g. 'hg38' for the human genome). Otherwise, custom TE coordinates can be provided as an additional input file in browser extensible data (BED) format (30). Unwanted tandem repeats and repetitive RNA classes are excluded (i.e. Low_complexity, Simple_repeat, rRNA, scRNA, srpRNA, tRNA). A whitelist can be provided to filter out invalid cell barcodes and speed up the workflow; for example, STARsolo's or Cell Ranger's filtered barcodes (i.e. the 'barcodes.tsv' file). The output of IRescue is a sparse matrix written in a Market Exchange Format file (MEX), compliant with the 10x Genomics Cell Ranger's output (27) to ensure compatibility with most toolkits for downstream analysis (31,32).

Mapping alignments on TEs

The first step of IRescue's workflow is to map the aligned reads to the TE genomic coordinates. Read alignments and TE coordinates are processed in parallel by chromosome, up to the number of allocated CPUs. The intersection between read and TE coordinates is performed by IRescue wrapping bedtools (33), to take into account eventual deletions and splitting events due to splice junctions (i.e. TE loci localized between donor and acceptor splicing coordinates are not considered mapped), increasing the mapping precision.

UMI deduplication algorithm

Reads carrying the same UMI sequence and mapped on the same set of TEs are used to build Equivalence Classes objects (ECs), preserving the read mappability information (i.e. whether the reads were uniquely mapped or multi-mapped):

$$EC = \begin{pmatrix} UMI\ sequence \\ Read\ IDs \\ TE\ names \end{pmatrix}$$

For each processed cell barcode, a directed graph is built where nodes represent ECs and directed edges represent potential PCR duplication events, connecting the template to the duplicate UMI. In order for two ECs to be connected, the UMI sequences must display up to 1 hamming distance (i.e. mismatches), the frequency (or read count) of the template's UMI must be at least double minus one the frequency of the duplicated UMI and at least one TE must be mapped by both. Formally, a directed edge connects node A to node B if the following conditions are satisfied:

- $hamming(A_{UMI},\ B_{UMI}) \leq 1$
- $A_{freq} \geq 2 \times B_{freq} - 1$
- $|A_{TEs} \cap B_{TEs}| \geq 1$

After building the cell-wide graph, the algorithm finds all subgraphs of connected nodes and, for each subgraph, calculates the deduplicated UMI count as following:

1. One or more nodes that do not receive an edge from other nodes (i.e. that are not PCR duplicates) are designated as *parents*.
2. A *pathfinder* function finds the paths starting from *parent* nodes through nodes sharing at least one TE, with each node assigned to exactly one path. Multiple configurations of paths can be found in a given subgraph.
3. The minimum number of paths found in a given subgraph corresponds to the subgraph's deduplicated UMI count, which is added to the respective TE counts (i.e. the TE contained in the *parent* nodes).
4. If a *parent* node contain more than one TE (i.e. corresponds to an EC generated by multi-mapping reads), the multi-mapped TEs are added to a cell-wide compatibility matrix, a logical matrix where rows are TEs, columns are UMIs and values of 1 or zero corresponds to the TE being mapped or not by each UMI. This matrix is used as input for the probabilistic redistribution of the multi-mapping UMI count in the step below.

Expectation-maximization

The cell-wide compatibility matrix then is used as a prior to redistribute the relative abundance of multi-mapping UMIs to the corresponding TEs by an Expectation-Maximization (EM) procedure. The EM consists in two steps repeated in several iterations: expectation (E) and maximization (M). In the E-step, the relative abundance of UMIs to each TE is optimized by dividing it for the running sum of the counts of the TEs mapped by the UMI:

$$R_{ij} = \frac{M_{ij}}{\sum_{k=1} E_{jk}}$$

Where M is a $m \times n$ matrix of the prior relative abundances of m UMIs across n TEs, R is a $m \times n$ matrix of the optimized relative abundances of UMIs across TEs, i and j indicate the ith row and jth column of the element of each matrix, E_j is an array containing the counts of TEs mapped by the UMI in the jth column. The M-step takes R as input and returns an array of optimized TE counts E by dividing each jth TE's abundance by the total UMI count in R:

$$E_j = \frac{\sum_{k=1} R_{kj}}{\sum_{i=1} \sum_{j=1} R_{ij}}$$

The EM procedure recurs until convergence or for a fixed number of iterations. Finally, the optimized counts from multi-mapped UMIs are summed to the uniquely mapped UMI counts to obtain the cell's final TE counts.

Writing final TE counts

The UMI deduplication and TE count estimation procedures are executed in multiple parallel processes, up to the number of allocated CPUs, to increase processing speed. Then, all the TE counts from all cells are written in a sparse matrix in the Market Exchange format (MEX), which is compliant with the output of Cell Ranger or STARsolo to ensure compatibility with several toolkits for single cell downstream analysis. IRescue is written in the Python programming language and leverages on essential open source libraries for efficient scientific computing and bioinformatics data wrangling (26,33–36).

Single-cell RNA sequencing data processing

Read alignments and single cell gene counts matrix were obtained by mapping scRNA-seq reads to the reference human genome (UCSC hg38 primary assembly) using STARsolo 2.7.9a (28,29). As genes and splice junction database, we used the Gencode comprehensive annotation v40 (37) with chromosome names converted to the UCSC nomenclature (passed to STAR through the '`--sjdbGTFfile`' parameter). We used the 10x Genomics cell barcodes whitelist for the v2 library kit (passed through the '`--soloCBwhitelist`' parameter), and the EmptyDrops algorithm (38) to filter out invalid cell barcodes. Other non-default parameters were: '`--outSAMattributes NH HI AS nM NM MD jM jI XS MC ch cN CR CY UR UY GX GN CB UB sM sS sQ --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --soloType CB_UMI_Simple --soloCellFilter EmptyDrops_CR 10000 0.99 10 45 000 90 000 500 0.01 20 000 0.01 10 000`'. For the 5′ PE dataset of human PBMCs in SARS-CoV-2 infection, additional parameters were used: '`--soloStrand Forward --soloBarcodeMate 1 --clip5pNbases 39 0 --soloUMIdedup 1MM_CR --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts`'. For single-nuclei data, the parameter '`--soloUMIlen 12`' was added. To estimate the expression of TEs at single cell level, read alignments were processed using IRescue 1.1.0, using the filtered barcodes list produced by STARsolo as cell whitelist (passed through the '`--whitelist`' parameter), and other parameters '`--genome hg38 --CBtag CB --UMItag UR --keeptmp --ec-dump`'. To compute TE counts with scTE 1.0 (19), the same Repeatmasker and Gencode annotations were used to build an index, keeping other parameters as default, and read alignments were counted with parameters '`-CB CB -UMI UR`'. To compute TE counts with SoloTE 1.09 (20), we generated the TE annotation compatible with SoloTE as per author's documentation and ran the quantification with default parameters.

Data simulations and benchmarks

Simulated scRNA-seq reads, UMI sequences and TE subfamily counts were obtained adapting the method in Kaminow et al. (28) for TE subfamily expression. This simulation procedure has the advantage to reproduce the 3′- or 5′-end positioning bias of droplet-based scRNA-seq reads aligning on any genomic region by using a real dataset as template. For this purpose, we used three 10x Genomics human PBMC datasets: 3′-end single-cell (https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0), 5′-end single-cell (https://www.10xgenomics.com/datasets/human-pbmc-from-a-healthy-donor-10-k-cells-v-2-2-standard-4-0-0) and 3′-end single-nuclei (https://www.10xgenomics.com/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-x-1-standard-2-0-0), which are common library layouts. Briefly, the simulated reads were derived from the alignment coordinates of real reads on a reference, along with the associated UMI sequences. In detail, cell barcodes are filtered according to the 10x Genomics whitelist and UMIs with uncalled bases are removed. Reads are aligned on a reference that combines the human hg38 primary genome assembly and the Repeatmasker TE genomic sequences using BWA-MEM 0.7.17 (39). UMIs are counted based on the TE sequence they map on; in case of alignments on multiple features, the top-scoring alignment is chosen. Finally, the sequences from alignment coordinates on the Repeatmasker reference were extracted. A mismatch rate of 0.5% was added in the extracted sequences to simulate Illumina sequencing errors and written in fastq format. Data for the plots showing the UMI deduplication statistics were obtained from the equivalence classes summary file generated by IRescue using the '`--dump-ec`' parameter. Spearman's correlation coefficients were calculated using R 4.3.1 between true and estimated TE counts in a pairwise manner across all the simulated library types. The statistical significance of fold changes between estimates and true counts for each TE subfamily was calculated on absolute values of $\log_2$(estimate/truth) to be able to compare both over- and under-estimated counts, using a two-tailed Wilcoxon rank sum test and P-values were adjusted for multiple comparisons as per Benjiamini-Hocheberg. Cell clustering according to TE counts was performed using the Louvain algorithm implemented in Seurat 5.0.1 (40), using the first 10 PCA dimensions with resolution 1.0. The cluster similarity index between measured and simulated clusters was calculated for each cluster using kBET (41) by transforming kBET's rejection rate to acceptance rate, as in Büttner et al. (41): $AcceptanceRate = 1 - RejectionRate$. Memory (RAM) usage and run time were measured by running the TE quantification methods in a Nextflow pipeline (42) using 1, 2, 4, 8 or 16 CPUs and extracted from the pipeline's tracing file.

Colorectal cancer scRNA-seq analysis

Data was obtained from ArrayExpress E-MTAB-8410. Data pre-processing and TE quantification were done as above using IRescue, scTE and SoloTE. TE counts normalization, principal component analysis (PCA), cell clustering (Louvain) and uniform manifold approximation and projection (UMAP) were performed using the Seurat 5.0.1 toolkit. Cell clusters based on TE expression profiling were annotated as tumoral (K) or normal (N) according to the most prevalent cell condition on each cluster. TE expression signatures were obtained by finding differentially expressed TEs across clusters using Seurat's FindAllMarkers function, and TEs significantly overexpressed (Wilcoxon rank sum test's P-value adjusted according to Bonferroni <0.01; $\log_2$ fold change > 1.5) on K or N clusters only were selected, discarding TEs overexpressed in clusters from both conditions. The difference in enrichment of TE subfamilies by class between tumor and normal signatures was tested with a two-tailed two-proportions Z-test using the R 4.3.1 stat package. The average expression of significantly overexpressed CRC marker TEs across clusters was visualized using Seurat's DotPlot visualization. Sashimi plots were obtained by loading the BAM files the IGV genome browser (43).

Human brain snRNA-seq analysis

Data was obtained from NCBI's Gene Expression Omnibus GSE209552. Data pre-processing, TE and gene expression quantification, normalization and scaling were done as above. Clusters based on gene expression were inferred with the Louvain algorithm with resolution 0.1, and annotated by cell type based on the expression of marker genes (44) as neurons (RBFOX3 for excitatory and GAD1 for inhibitory neurons) or glia (GFAP for astrocytes, PLP1 for oligodendrocytes, VCAN for OPCs, FYB1 for microglia). Clusters based on TE expression were inferred with 0.5 resolution and annotated as neuronal or glial based on the prevalent cell type. TE subfamilies specific for each cluster were identified as above, with adjusted P-value <0.05 and average $\log_2$ fold change >0.5.

Human skin aging scRNA-seq analysis

Data was obtained from NCBI's Gene Expression Omnibus GSE130973. Data pre-processing and TE quantification were done as above. TE counts were normalized with gene's counts to accurately estimate the per-cell library size as $log1p(\frac{Count}{Cell\ libSize} \times 10,000)$ using Seurat's NormalizeData function. Data was scaled and centred with Seurat's ScaleData function using all TE subfamilies. Next, for gene and TE counts, PCA was calculated, batches integration was done by canonical correlation analysis (CCA) using the first 20 PCA dimensions and UMAP using the first 20 PCA dimensions for genes and 10 for TEs. To identify cell types, cells were clustered based on gene expression at 0.7 resolution and the following marker genes were used, as in (45): LYZ, AIF1, HLA-DRA, CD68, ITGAX (macrophages and dendritic cells); CD3D, CD3G, CD3E, LCK (T cells); SELE, CLDN5, VWF, CDH5 (vascular endothelial cells); LYVE1, PROX1 (lymphatic endothelial cells); ACTA2, RGS5, PDGFRB (pericytes); HBA1, HBA2, HBB (erythrocytes); PMEL, MLANA, TYRP1, DCT (melanocytes); KRT5, KRT14, TP63, ITGB1, ITGA6 (undifferentiated keratinocytes); KRT1, KRT10, SBSN, KRTDAP (differentiated keratinocytes); LUM, DCN, VIM, PDGFRA, COL1A2 (fibroblasts). For fibroblasts-derived cell subsets: CCN5, SLPI, CTHRC1, MFAP5, TSPAN8 (secretory-reticular); APCDD1, ID1, WIF1, COL18A1, PTGDS (secretory-papillary); CCL19, APOE, CXCL2, CXCL3, EFEMP1 (pro-inflammatory); ASPN, POSTN, GPC3, TNN, SFRP1 (mesenchymal). For T cell subsets: CCR7, TCF7, LEF1, SELL (Naïve); CD27, CD28, PTPRC, IL7R (Memory); IL2RA, IFNG, IL7R-, (Effector); NKG7, GNLY, GZMH, GZMB, CCL4 (Cytotoxic lymphocytes, CTL). Genes or TEs specific for each fibroblasts or T cell subpopulations were identified with Seurat's FindAllMarkers function as above. Average TE or gene expression per cell type across donor's age were extracted with Seurat's AverageExpression function and plotted using the R package pheatmap (https://github.com/raivokolde/pheatmap). TEs differentially expressed between elderly and adults were identified using Seurat's FindMarkers function (adjusted P-value < 0.05 and average log2 fold change > 0.5 or < -0.5) and visualized in volcano plots using the R package ggplot2 (https://ggplot2.tidyverse.org). Enriched TE subfamilies in each TE class across cell types were calculated as above, dividing between TE subfamilies upregulated or downregulated in aging.

PBMCs in SARS-CoV-2 infection scRNA-seq analysis

Data was obtained from ArrayExpress E-MTAB-9652. Data normalization, scaling and dimensionality reduction were done as for human skin aging, with the exception of using Reciprocal PCA (RPCA) instead of CCA for integration due to the high size of the dataset. TE average expression across groups were extracted and plotted as for human skin aging, with TE family-level expression being calculated by summing TE subfamilies counts by family. TE differential expression, TE subfamilies enrichment, heatmaps and volcano plots were done as for human skin aging, and feature selection were done by adjusted P-value < 0.05 and average $\log_2$ fold change >0.25 or <-0.25.

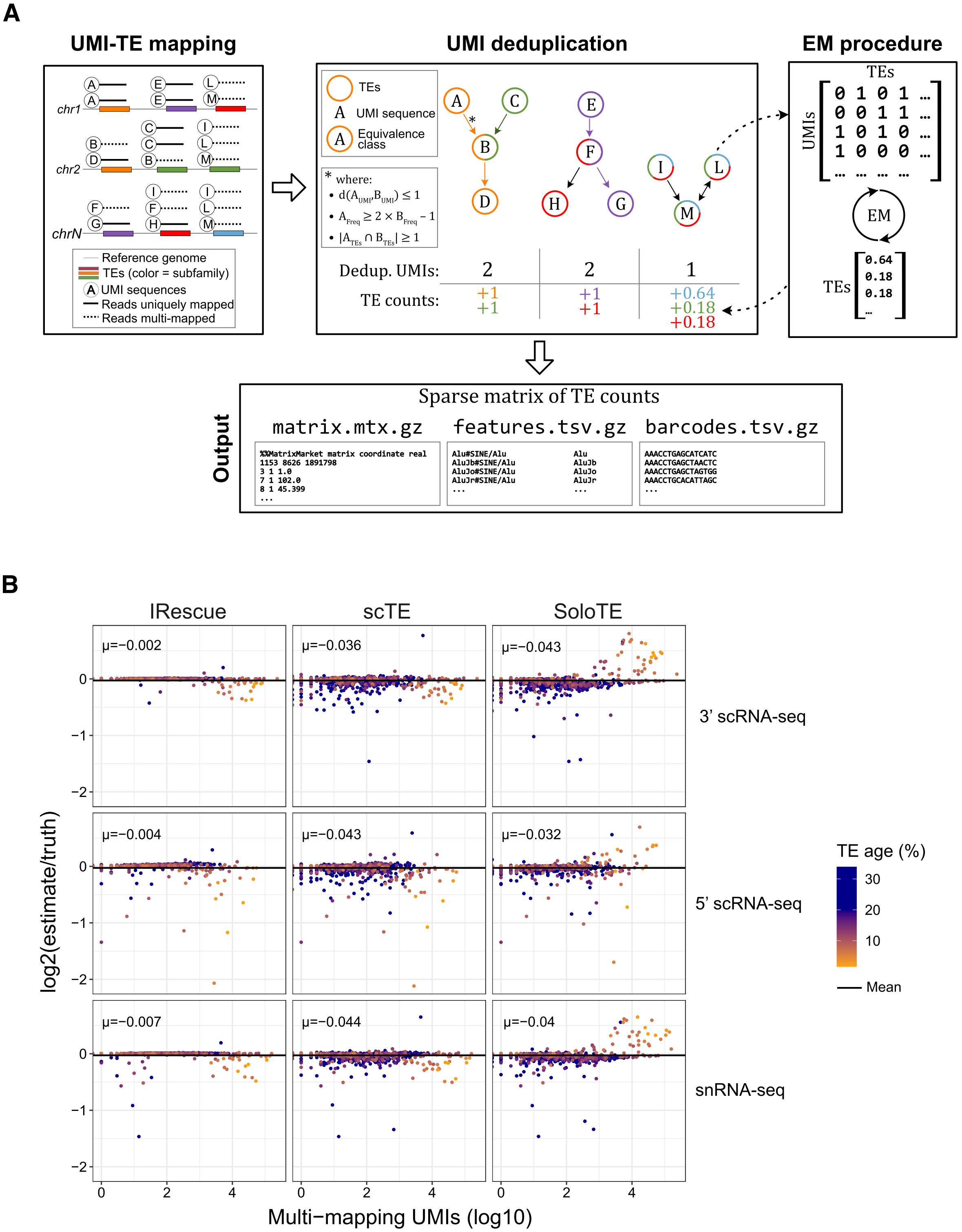
Results

IRescue: a novel multi-mapping aware algorithm for the quantification of TE expression in scRNA-seq data

Rescuing multi-mapping reads is widely recognized as an essential step to precisely quantify the expression of TEs (5) or other challenging-to-map features, like multigene families (46). However, currently published computational tools for the quantification of TE subfamilies expression at single-cell level (19,20) select only one random alignment of multi-mapped reads and do not account for eventual sequencing errors in the UMI sequences, without implementing state-of-the-art approaches for UMI deduplication and multi-mapping reads quantification. These shortcomings impair the correct estimation of many TE subfamilies, in particular the young ones. Hence, we have developed IRescue, an algorithm that is able to allocate the counts from multi-mapped reads in a probabilistic manner, while also accounting for sequencing errors in the UMI sequence and PCR duplicates. To accomplish this, we modified a state-of-the-art UMI deduplication procedure (23), adjusting it for UMIs mapped across interspersed repeats instead of fixed genomic coordinates. For the first time, we have introduced an Expectation-Maximization (EM) algorithm to probabilistically redistribute the signal from all multi-mapping UMIs that remain associated to more than one TE subfamily after deduplication. Briefly, reads sharing the same UMI sequence and mapped feature are represented by a single object called equivalence class (EC). ECs are organized in a cell-wide directed graph, where direct edges connect the original UMIs with their PCR duplicates. Subsequently, the deduplicated UMI count is found by calculating the minimum number of paths in each subgraph of connected nodes. In cases where multiple TE subfamilies are linked to the graph, the UMI count is optimally redistributed among subfamilies through the EM convergence (Figure 1A). The TE expression estimates represent three components: gene-TE chimeric transcripts, pervasive transcription and transcription of entire TE elements (5). Only few young TE insertions contribute to the latter in human, whereas older elements can only be transcribed within other transcriptional units. Since different UMI-based single-cell technologies capture different portions of transcripts, we assessed the accuracy of IRescue by simulating both 3′-end and 5′-end scRNA-seq datasets, as well as single-nuclei RNA-seq (snRNA-seq). We observed that, using IRescue, UMIs are deduplicated by about 4-fold (Supplementary Figure S1A) and, following deduplication, the amount of ambiguous UMIs associated to multiple TE subfamilies was drastically reduced (Supplementary Figure S1B), showing how a sensible UMI-deduplication procedure can reduce noise in single-cell TE analysis. Next, we tested the performance of IRescue and alternative tools, scTE (19) and SoloTE (20). Overall, IRescue's counts better correlated with their simulated counterpart, with only minor differences between tools observed in the 5′-end dataset (Supplementary Figure S1C-E). We attributed this discrepancy to the significantly lower presence of TEs in the 5′-end of transcripts (47), which reduces the amount of ambiguous alignments and, consequently, the benefit of using IRescue's algorithm. As expected, the correlation between different single-cell library types, whether using simulated or estimated mean TE expression, was low, confirming that library composition has a strong impact on TE detection (24,48) (Supplementary Figure S1F). We evaluated the quantification precision of IRescue and alternative tools considering both TE age and the frequency of reads aligning to multiple locations, which are more common on younger TEs (12) (Supplementary Figure S1G). Our findings demonstrate that IRescue significantly improves the quantification precision of TE subfamilies with extensive read alignment to multiple genomic positions (Figure 1B), and showed better performance for old TEs as well (Supplementary Figure S1H). Notably, a small number of TE subfamilies (9 in 5′ scRNA-seq and 4 in snRNA-seq) were significantly underestimated by IRescue (less than –0.5 $\log_2$ fold change) and consistently underestimated by scTE and SoloTE as well, suggesting an issue in the mappability of these TEs during read alignment rather than in quantification (Figure 1B, Supplementary Table S1). Next, we assessed the performance of IRescue, scTE and SoloTE in inferring cell clusters using TE expression and observed a higher cluster similarity between true and inferred clusters when employing IRescue in 3′ scRNA-seq and snRNA-seq, with no significant difference among them in 5′ scRNA-seq (Supplementary Figure S1I). Finally, we demonstrated that IRescue utilizes computational resources more efficiently than other tools, while maintaining a reasonable runtime (Supplementary Figure S1J). Overall, we demonstrated that IRescue outperforms existing tools in the accuracy of TE expression quantification in UMI-based single cell datasets, especially for 3′ scRNA-seq and snRNA-seq.

IRescue enables the identification of tumor-specific TE signatures in colorectal cancer at single-cell resolution

To evaluate the performance of IRescue against scTE (19) and SoloTE (20) in real datasets, first we leveraged the well-characterized expression patterns of various transposable element (TE) subfamilies in colorectal cancer (CRC) (17,49,50). We analysed the TE expression profiles in a publicly available 10x Genomics 3′ scRNA-seq dataset of tumor and adjacent normal tissues from six CRC patients (51). We clustered the cells based on TE expression and visualized them on a 2D UMAP space, identifying five clusters predominantly or exclusively composed of cancer cells (K1–5) and six by normal cells (N1–6) (Figure 2A, B), confirming that TE expression enables the discrimination between cancerous and normal cells within the dataset. In particular, we found 75 TE subfamilies overexpressed specifically in cancer cell clusters and 39 in normal cell clusters. The expression of more TE subfamilies is detectable in a larger fraction of cells belonging to cancer clusters, compared to normal ones (Supplementary Figure S2A and B, Supplementary Table S2). LINE and LTR subfamilies were significantly over-represented in cancer, in accordance with literature (52,53), whereas SINE subfamilies abundance did not change significantly between the two conditions (Figure 2C). Interestingly, among LINE subfamilies, those that were differentially expressed in normal cells were evolutionarily old LINE1 (i.e. L1M*, mammalian-wide LINE1 elements), whereas younger LINE1 were specific for cancer cells only (i.e. L1HS, L1PA* and L1P*, human and primate-specific) (Figure 2D). Also scTE and SoloTE tools identified distinct cancer cell clusters characterized by a general enrichment of LINE and LTR subfamilies (Supplementary Figure S2D-F). We checked the expression of 16 TE subfamilies known to be overexpressed in CRC (17,49) across the cell clusters identified in scRNA-seq and confirmed the presence of eight subfamilies belonging to the cancer TE signature using IRescue (Figure 2E and Supplementary Figure S2C). Despite L1HS and L1PA2 being recognized as overexpressed in CRC (17), scTE failed to detect them as differentially expressed and SoloTE identified them as expressed indiscriminately in both cancerous and normal cells (Figure 2E, Supplementary Figure S2G). This highlights that IRescue, by virtue of its improved quantification of multi-mapping TE reads, enhances the detection of CRC-specific young TE subfamilies that are otherwise challenging to pinpoint in single-cell RNA-seq data. Given previous reports of L1PA2-SYT1 and MER1B-PIWIL1 as distinctive alternative isoforms formed by TE insertions in CRC (50), we inspected the alignment of reads across these genomic loci. We sought evidence of reads spanning splice junctions mapped onto both TEs and non-repetitive genomic regions, indicative of potential hidden TE-containing exons. We found over 3-fold more reads splitted between the L1PA2-derived exon and the third exon of SYT1 and confirm the detection



```
%%MatrixMarket matrix coordinate real
1153 8626 1891798
3 1 1.0
7 1 102.0
8 1 45.399
...
```

```
Alu#SINE/Alu                 Alu
AluJb#SINE/Alu               AluJb
AluJo#SINE/Alu               AluJo
AluJr#SINE/Alu               AluJr
...                          ...
```

```
AAACCTGAGCATCATC
AAACCTGAGCTAACTC
AAACCTGAGCTAGTGG
AAACCTGCACATTAGC
...
```

Figure 1. IRescue's algorithm schematics and benchmarking. (**A**) Scheme of IRescue's algorithm. (Top-left) IRescue takes as input uniquely mapped and multi-mapped reads aligned on a reference genome with annotated UMI and barcode sequences in BAM format. (Top-middle) equivalence classes (ECs) containing each UMI's sequence, frequency and mapped TE subfamilies are used to build a directed graph according to the indicated conditions (i.e. minimum hamming distance, frequency difference and TE subfamilies in common). For each subgraph of connected nodes, the minimum number of paths is calculated to obtain the deduplicated UMI count, which is assigned to the corresponding TE subfamily. (Top-right) In case of deduplicated UMIs being associated to more than one TE subfamily, an Expectation-Maximization (EM) procedure redistributes the UMI's relative abundance to optimize the expression estimate of each subfamily. (Bottom) the UMI counts per TE subfamilies in each cell are written in a Matrix Market exchange format (Cell Ranger-compatible) for downstream analysis. (**B**) Scatterplots of TE subfamilies (*N* = 1202) showing the relationship between the number of the associated multi-mapping UMIs and the fold change between estimated and true counts in the indicated dataset and quantification method. Each dot represents a TE subfamily, color-coded by the average insertion age presented as the percentage of divergence between genomic TEs and their respective consensus sequence (as reported in UCSC's Repeatmasker annotation). Blue dots correspond to older TEs, whereas orange dots represent younger TEs. Black horizontal lines and μ indicate the mean.

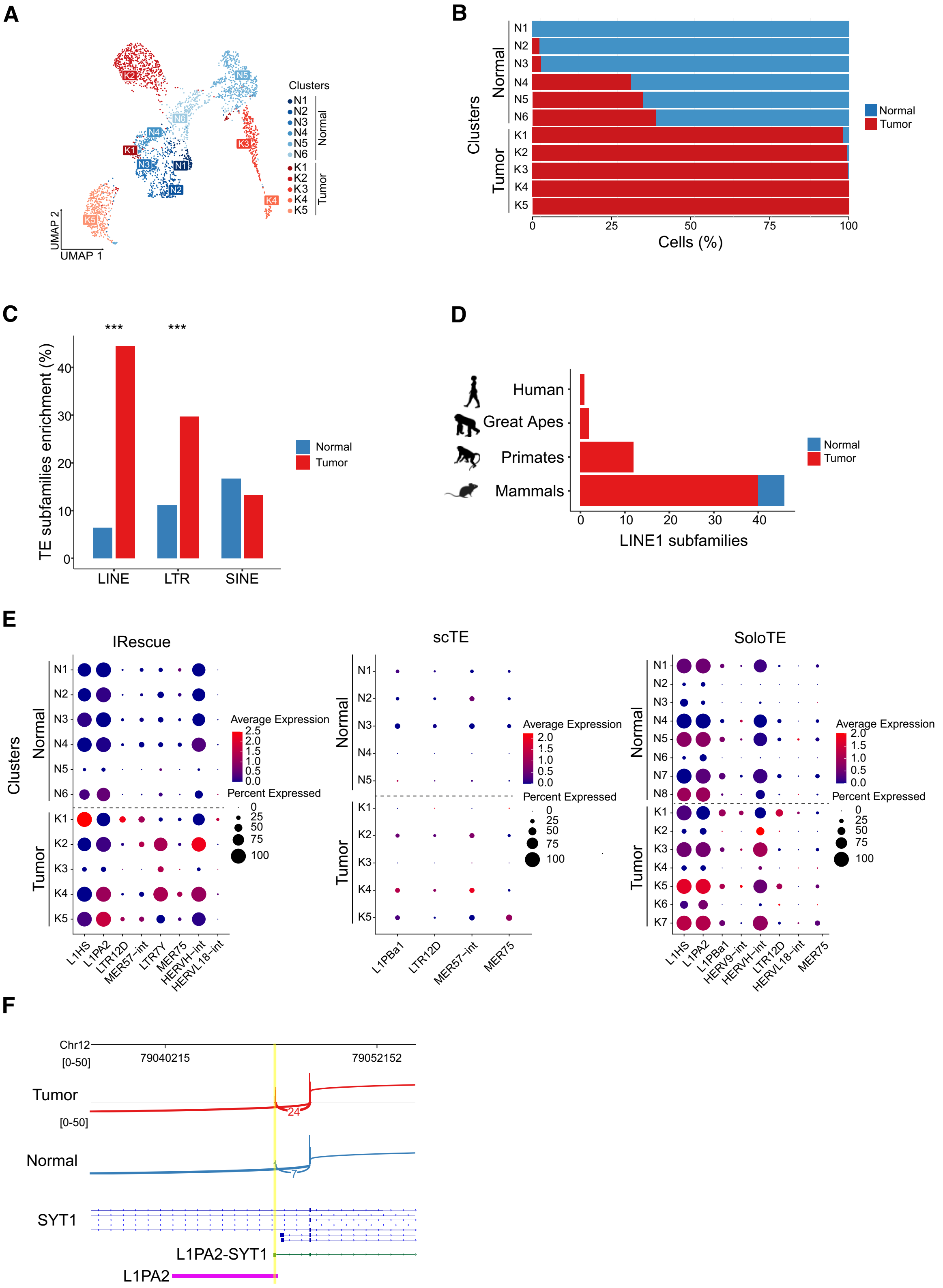


Figure 2. Identification of TE expression dynamics in colorectal cancer. (**A**) UMAP representation of CRC and normal cells according to TE expression. Clusters of normal and cancer cells (indicated in legend) are obtained on the basis of TE expression. (**B**) Relative abundance of cells by condition across clusters. (**C**) Enrichment of differentially expressed TE subfamilies in normal or cancer condition (adjusted *P*-value < 0.05) by TE class, calculated as the percentage in respect to the total number of subfamilies per class. *** *P*-value < 0.001 (two-sided two-proportions *Z*-test). (**D**) Number of differentially expressed LINE1 subfamilies in normal or cancer condition (adjusted *P*-value < 0.05) by evolutionary clade (Human: L1HS; Great apes: L1PA[2–3]; Primates: L1P*; Mammals: L1M*). Animal shapes were obtained from PhyloPic and are copyright-free. (**E**) Average expression of differentially expressed known TE CRC markers across clusters using the indicated quantification method. The dot size is indicative of the percentage of expressing cells in the cluster (adjusted *P*-value < 0.05). (**F**) Sashimi plot representing the coverage across the splice junction of a L1PA2-derived CRC-specific alternative cryptic exon of the SYT1 oncogene.

of a MER1B-PIWIL1 TE-oncogene isoform in CRC (Figure 2F and Supplementary Figure S2H), concluding that IRescue correctly estimates the expression of TEs transcribed within specific TE-containing transcripts.

IRescue dissects the expression dynamics of old and young LINE1s in single nuclei human brain

To test IRescue and alternative tools' capability to quantify TE subfamily expression in single nuclei, we analysed the TE expression dynamics in human brain, a biological context known to be characterized by clear LINE1 activity (54–56) and expression patterns (44) in neurons. We processed a snRNA-seq dataset of temporal and frontal lobes from five adult donors (44). After identifying the main neuronal and glial populations using the expression of marker genes, we performed clustering and dimensionality reductions based on TE subfamily expression estimates with IRescue, revealing a clear separation between neuronal and glial nuclei. In particular, six clusters were mostly composed by neuronal and two clusters by glial nuclei (Figure 3A, B). This was further confirmed by performing the same analysis with scTE and SoloTE, which led to similar results (Supplementary Figure S3A, B). It has been reported that LINE1 subfamilies are expressed in adult human neurons, but not in glia, at bulk or pseudo-bulk level (44). Here, we investigated whether specific LINE1 subfamilies were enriched in distinct neurons clusters at single-nucleus resolution. We detected 35 LINE1 subfamilies significantly overexpressed in neurons clusters, and none in glia (Figure 3C). Interestingly, the expression of young LINE1 subfamilies (L1HS and L1PA2-4) were mostly restricted to one specific neurons cluster (Figure 3C, D), whereas most evolutionarily old LINE1s were expressed in multiple clusters, different from the cluster enriched in young LINE1 expression (Figure 3C, E). Finally, we tested the detection of differentially expressed LINE1 subfamilies between nuclei clusters using scTE and SoloTE. With the former method, we found 27 LINE1 subfamilies specific to neuronal clusters and none in glial clusters. Conversely, using SoloTE, we detected 28 LINE1 subfamilies specific to neuronal clusters and 4 to glial clusters (Figure 3F, Supplementary Figure S3C, Supplementary Table S3). This was consistent to our previous findings on both simulated and real data which indicated that scTE slightly underestimate TE expression while SoloTE overestimates it, with IRescue representing an improvement over both methods.

IRescue unveils the varying patterns of retrotransposon expression during human skin aging at the single-cell level

As an additional validation for IRescue, we inspected the TE expression dynamics in human aging, during which TEs are upregulated due to epigenetic de-repression and structural changes in the genome (57–61). The dynamics of TEs in the context of human aging remain inadequately explored, particularly at the single-cell level. Therefore, we analyzed a 3′ scRNA-seq dataset comprising 20676 cells collected from five male donors aged between 25 and 70 years old, with a focus on human skin aging (45). By assessing the expression of marker genes, we delineated 10 distinct cell types (Supplementary Figure S4A) and, when employing dimensionality reduction based solely on TE expression, we were able to distinguish between most cell types, although with slightly lower precision (Supplementary Figure S4B). Firstly, we focused on the four primary subsets of dermal fibroblasts, which are recognized for experiencing substantial functional decline with age (45,62): secretory-papillary, secretory-reticular, mesenchymal and pro-inflammatory (Figure 4A). We explored the expression dynamics of genes and TE subfamilies significantly enriched in each fibroblast subset across all samples. When grouping samples by hierarchical clustering based on gene expression, cell types from donors of different ages clustered together (Supplementary Figure S4C). In contrast, applying the same clustering procedure based on TE expression resulted in the grouping of different cell types by donor age, except for mesenchymal fibroblasts (Figure 4B). Notably, elderly individuals (aged 69–70 years) exhibited an over-expression of SINE and LINE elements compared to younger adults, suggesting a shift in TE expression dynamics in aging (Figure 4B). We then identified the specific TE subfamilies exhibiting differential expression between elderly (aged 69–70 years) and adult (aged 25–53 years) individuals. We found an overall upregulation of LINE1 and Alu subfamilies in elderly fibroblasts, consistent across all fibroblast subsets except for mesenchymal cells, while most ERV subfamilies showed downregulation (Figure 4C,D and Supplementary Table S4). It is noteworthy that mesenchymal cells are reported to undergo fewer functional declines during aging compared to other fibroblast subsets (45). These results imply a correlation between the expression dynamics of specific TE families and fibroblast functions, with mesenchymal cells showing relatively minor alterations in TE expression profiles. The most recognized transcriptional pattern of TEs in aging involve the activation of specific TEs that are normally repressed in young individuals due to loss of DNA methylation. For this purpose, we assessed the differential expression of Alu, LINE1 and ERV1 subfamilies that are reported to follow this pattern (57,63,64) using IRescue, scTE and SoloTE. All the 12 analysed TE subfamilies are enriched in elderly individuals using IRescue's estimates, with six being significantly upregulated (Figure 4E). A similar result is observed using scTE and SoloTE, with the exception of AluYb9, a subfamily that varies from AluYb8 by only a single nucleotide replacement (65), which in contrast was found to be more expressed in adults in both the tools (Figure 4E). Next, we analysed the 2634 T cells present in the same human aging skin dataset described above (45). It has been previously documented that the dynamics of TEs varies across different T cell functional states (9), but it remains unclear whether aging affects TE expression dynamics in T cells. The expression of marker genes identified four main T cell type subsets: naïve, memory, effector and cytotoxic (Supplementary Figure S4D, E). However, TE expression profiles clustered these subsets differently according to age, with a minor exception observed in naïve T cells (Supplementary Figure S4F). Notably, elderly individuals (aged 69–70 years) exhibited an overexpression of LINEs and SINEs compared to younger adults (Supplementary Figure S4F), suggesting that the increased expression of these two classes of retrotransposons is a common characteristic of cellular aging across various cell types.

IRescue reveals varied TE expression in the innate and adaptive immune systems throughout SARS-CoV-2 infection and subsequent recovery.

To further showcase the effectiveness of IRescue in accurately quantifying the expression of TE subfamilies in single cells,

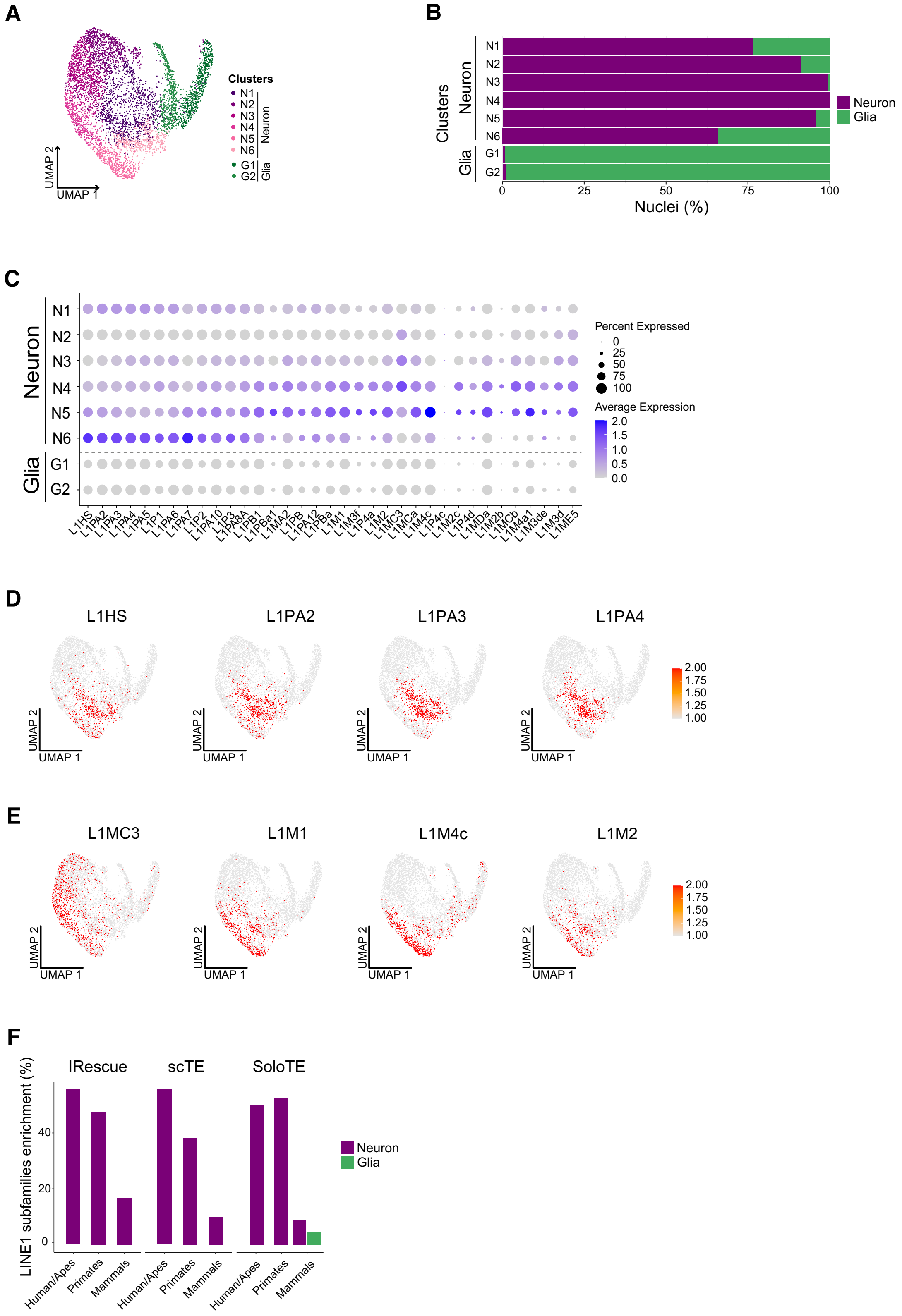


Figure 3. LINE1 are dynamically expressed in the human brain in specific subpopulations of neuronal nuclei. (**A**) UMAP representation of neuronal and glial nuclei according to TE expression, colored by cluster identity inferred by TE expression and cell type inferred by gene expression. (**B**) Relative abundance of nuclei by major cell type across TE clusters. (**C**) Average expression of differentially expressed LINE1 subfamilies across clusters. The dot size is indicative of the percentage of expressing cells in the cluster (adjusted *P*-value < 0.05, average $\log_2$ fold change > 0.5). (**D**) UMAP representation of nuclei colored by scaled expression of evolutionarily young LINE1 subfamilies. (**E**) as (D), for evolutionarily old LINE1 subfamilies. (**F**) Enrichment of LINE1 subfamilies differentially expressed in neurons or glia among the total number of annotated subfamilies, stratified between L1HS/PA (human or apes), L1P (primates) and L1M (mammals), using the indicated quantification method.

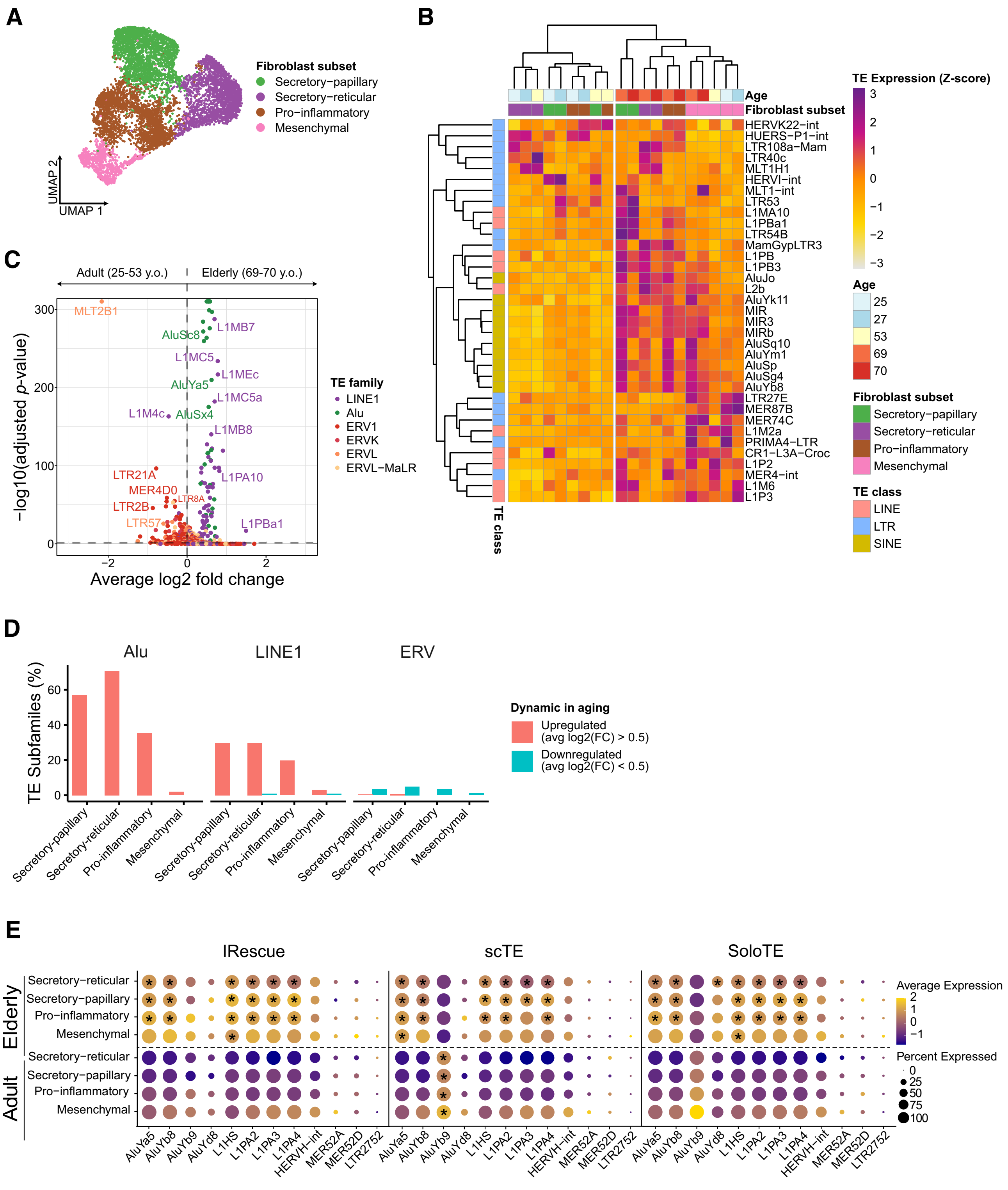


Figure 4. Human skin fibroblasts and T cells display specific single-cell TE expression patterns in aging. (**A**) UMAP representation of skin-derived fibroblasts subsets according to gene expression. (**B**) Average expression of TE subfamilies differentially expressed in each fibroblast subset, further stratified by donor's age (top 10 significant TE subfamilies per cell type, adjusted *P*-value < 0.05). Expression values are normalized and scaled by *Z*-score. Dendrograms display the hierarchical clustering of TE subfamilies (rows) and samples (columns) according to TE expression patterns. The color code indicates the TE class (rows), donor's age or fibroblast subset (columns). (**C**) Volcano plot of TE subfamilies by average $\log_2$ fold-change between elderly and adults individuals and adjusted *P*-value in negative $\log_{10}$ scale, colored by TE family. Horizontal dashed line indicates *P*-value = 0.05, vertical dashed line indicates $\log_2$ fold change = 0. (**D**) Enrichment of differentially expressed TE subfamilies (adjusted *P*-value < 0.05) in aging by fibroblasts subsets and TE families, calculated as the percentage in respect to the total number of subfamilies per family. ERV family includes ERV1, ERVK, ERVL and ERVL-MaLR. (**E**) Average expression of TE subfamilies expected to be upregulated in aging across fibroblasts subsets using different quantification methods. The dot size is indicative of the percentage of expressing cells in the cluster. * indicates that the TE subfamily is significantly upregulated in adult or elderly condition (adjusted *P*-value < 0.05, average $\log_2$ fold change > 0.5).

we investigated a recent case study involving TE upregulation in human peripheral blood mononuclear cells (PBMCs) following recovery from SARS-CoV-2 infection. During the acute phase of infection, TE expression is reported to remain largely unchanged in PBMCs (66–68). However, strong TE upregulation has been documented in PBMCs after patients have recovered from COVID-19, with the extent of upregulation correlating with the severity of the disease (69). Hence, we employed IRescue to analyze a 5′ scRNA-seq dataset comprising 87324 PBMCs obtained from patients experiencing either mild or severe symptoms in acute and recovery phase (70). First, we globally evaluated the expression of the most abundant families of TEs (LINE1, Alu and ERVs) in all PBMCs, stratifying the samples according to disease's severity and infection state. As reported (66–69), we observed a subtle increase in TE expression during the acute phase of infection compared to controls and a marked upregulation during the recovery phase, (Supplementary Figure S5A). By analyzing the expression of lineage-specific marker genes, we identified four main populations: monocytes, natural killer (NK), T and B lymphocytes (Figure 5A). To obtain a granular insight on TE dynamics from the acute to the recovery state, we conducted TE differential expression analysis in each immune lineage comparing COVID-19 patients and healthy donors. Our findings revealed that all TE families exhibited increasing over-representation from acute to recovery state in T and B cells, with Alu subfamilies showing the highest increase during recovery. Interestingly, Alu displayed an opposite pattern in monocytes and NK cells. Moreover, while LINE1 and ERV subfamilies demonstrated significant increases in T and B lymphocytes during recovery, they showed a weaker upregulation in monocytes and NK cells (Figure 5B). Next, we explored the expression of the differentially expressed TEs according to the severity of the disease. We observed that healthy donors' monocytes and NK cells showed a strong TE expression signature, whereas T and B cells were better identified in severe patients (Figure 5C, Supplementary Figure S5B). Interestingly, healthy donors' monocytes and NK cells were enriched by Alu subfamilies, compared to both acute infection and remission samples (Figure 5D, Supplementary Figure S5C); in contrast, T and B cells showed a strong Alu upregulation specifically in COVID-19 remission, with a greater significance in severe patients (Figure 5E, Supplementary Figure S5D). Both young and old LINE1 subfamilies were among the upregulated ones in acute infection and remission in all lineages, whereas ERVs showed a more mixed behavior, since some specific ERV subfamilies defined the signature of healthy donors (e.g. MER21B, LTR8) and others were specific for infected patients (e.g. LTR16D1, LTR83) (Figure 5D-E, Supplementary Figure S5C-D, Supplementary Table S5). Importantly, the majority of the Alu and ERV subfamilies reported to be upregulated in SARS-CoV-2 recovery (69) were identified in our differential expression analysis in T and B lymphocytes, and were similarly enriched among all upregulated TE subfamilies by either IRescue or alternative tools (Supplementary Figure S5E, Supplementary Table S5). Overall, through the analysis of real single cell/nuclei RNA-seq datasets, we validated IRescue capability to dissect the expression patterns of TE subfamilies in different biological contexts where the expression behavior of TEs has been previously reported.

Discussion

Estimating transposable element (TE) expression poses a persistent computational challenge due to the nature of repetitive elements, where reads can align equally well to multiple genomic loci (5). Typically, aligners either report a random alignment per read or arbitrarily designate one alignment as primary and others as secondary, which can result in inaccurate TE estimates. While several tools for TE expression quantification have been developed for bulk RNA-seq to address secondary alignments when assigning multi-mapping reads to TEs (5,13), similar strategies have not yet been implemented in the single-cell field. In this work, we introduce IRescue, a novel approach to accurately estimate TE expression in scRNA-seq data. IRescue not only considers all the secondary alignments of multi-mapping reads for TE assignment, but also provides the first strategy for deduplicating the UMIs associated to such reads, taking into account UMI frequencies and sequencing errors. Additionally, IRescue employs an Expectation-Maximization algorithm to probabilistically redistribute the count of UMIs still associated with more than one TE after deduplication—a procedure commonly used for processing multi-mapping reads (71–73) and previously implemented only in bulk-level TE quantification (15,16,74). In details, through these implementations, we demonstrated that IRescue is the most precise tool in quantifying TE expression at subfamily level when compared to other state-of-the-art tools, scTE (19) and SoloTE (20), using simulated scRNA-seq 10x Genomics-like reads and real datasets from different type of libraries (5′ and 3′ scRNA-seq, snRNA-seq). Reads simulated from TE sequences are more prone to map to multiple genomic regions, as they only contain the repetitive sequence and no flanking sequences, and therefore it is crucial to evaluate the ability of TE quantification methods in correctly estimating their contribution to TE expression. In contrast, chimeric alignments are composed in part by the TE sequence and in part by non-repetitive genomic sequences, and therefore are more easily allocated to their respective TE subfamily's counts. In this context, it is important to note that simulated data might display fewer chimeric alignments compared to real data. For this reason, we tested IRescue and alternative tools in four different case studies using real experimental data. Using simulations, we showed that IRescue is more accurate in allocating and quantifying multi-mapping reads and this improvement concerns both evolutionarily old and young TE subfamilies, the latter being more prone to be multi-mapped. This feature makes IRescue more reliable in the identification of young elements like L1HS and L1PA2, known to be overexpressed in colorectal cancer (17,49). In fact, IRescue identified those TEs as enriched and heterogeneously expressed in distinct cancer cell clusters, while scTE and SoloTE failed to detect them or observed their expression in the same cancer clusters and, at lower levels, in normal cells as well. In snRNA-seq, IRescue detected more LINE1 subfamilies specific to neurons against glia in human cells, compared to scTE and SoloTE, and identified a distinct neurons cluster enriched in young LINE1 expression. IRescue's accuracy was also validated in the investigation of TE dynamics during aging, with an upregulation of LINE1 and Alu elements in elderly individuals (58,61) and the identification of TE subfamilies specifically expressed in different fibroblast subsets, such as AluSp, AluSg4, L1PBa1, L1MA10 and upregulated in aging due to methylation loss, such as AluYa5 and L1HS. Moreover,

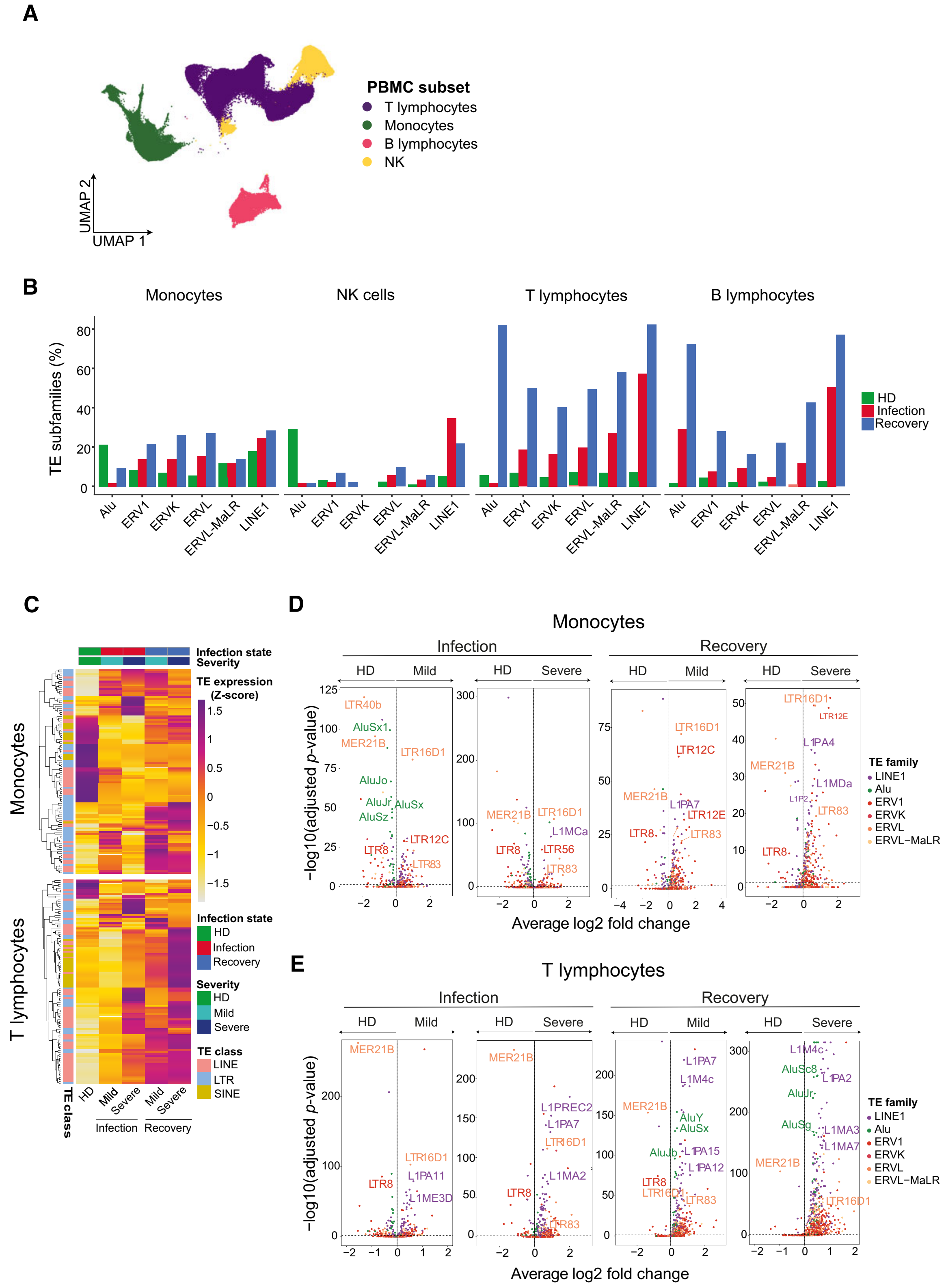


Figure 5. Specific TE expression patterns characterize human immune cells during SARS-CoV-2 infection and recovery. (**A**) UMAP representation of PBMCs colored by cell type inferred by gene expression. (**B**) Enrichment of differentially expressed TE subfamilies (adjusted *P*-value < 0.05, average $\log_2$ fold change > 0.25 or < −0.25) in infection or recovery compared to healthy conditions for the indicated cell types, calculated as the percentage in respect to the total number of TE subfamilies per family. (**C**) Average expression of differentially expressed TE subfamilies between healthy and infected or recovery conditions in the indicated cell type (adjusted *P*-value < 0.05, average $\log_2$ fold change > 0.25 or < −0.25), displaying the 50 most significant TEs per comparison. The color code indicates the TE class, disease's severity or patient's condition. Normalized expression is scaled by *Z*-score. Dendrograms display the hierarchical clustering of TE subfamilies according to expression patterns across groups. (**D**) Volcano plots of TE subfamilies by average $\log_2$ fold-change between the indicated conditions and adjusted *P*-value in negative $\log_{10}$ scale in monocytes, colored by TE family. Horizontal dashed line indicates *P*-value = 0.05, vertical dashed line indicates $\log_2$ fold change = 0. (**E**) As in (D), in T lymphocytes.

IRescue corroborated that no TE alteration occurs in the acute phase of SARS-CoV-2 infection, while a strong TE upregulation happen during recovery phase (66–69). Interestingly, IRescue detected most of ERV and Alu subfamilies reported in literature as differentially expressed in T and B lymphocytes (69), with a similar enrichment observed in alternative tools. Furthermore, we demonstrated that IRescue is more powerful in inferring cell clusters compared to alternatives, showing a higher similarity with simulated cells clusters and avoiding technical biases in defining the TE expression profile. Notably, in case of datasets with low complexity of TE signal, such as 5′ scRNA-seq (47), IRescue showed the least differences against scTE and SoloTE in both simulations and real data. This is expected, since the benefit of IRescue's algorithm resides in the optimized quantification of a large amount of ambiguous multi-mapping TE reads. IRescue enables sophisticated analysis of TEs in scRNA-seq datasets, highlighting the increasingly predominant role of TEs in regulating cellular functions; for instance, it accurately quantifies the signal from reads mapping on both TEs and unique genomic regions, which indicate the presence of chimeric transcripts (5). In this regard, we presented evidence of TE-originated transcript variants of human oncogenes in scRNA-seq (50,75), which incorporate TEs that, using IRescue, were detected as overexpressed in specific cancer cells clusters. We highlighted the presence of a small number of TE subfamilies consistently underestimated by all tested quantification methods, suggesting that this issue arises at the level of read alignment—a pre-requisite for quantification and common to all the three tools. Further investigations could enhance the state of alignment methods in scRNA-seq by identifying optimal parameters to improve the mapping for these problematic TEs. Overall, the transcription of TEs is involved in several physiological scenarios, such as embryo development (10), aging (76), differentiation and cell identity (9), as well as pathological conditions, such as cancer (52,77) and neurodegenerative diseases (78). However, the expression dynamics and heterogeneity of the single cell TE transcriptome in most of these tissues is still unexplored. With the release of IRescue, we long to facilitate the single cell TE expression profiling from canonical scRNA-seq experiments. This will open up the possibility of extracting novel insights from the vast amount of publicly available datasets (22).

Data availability

IRescue source code and documentation are available at https://github.com/bodegalab/irescue and https://doi.org/10.5281/zenodo.13479364. Data used for benchmarks is available at 10x Genomics (https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0, https://www.10xgenomics.com/datasets/human-pbmc-from-a-healthy-donor-10-k-cells-v-2-2-standard-4-0-0, https://www.10xgenomics.com/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-x-1-standard-2-0-0), colorectal cancer scRNA-seq data at ArrayExpress (E-MTAB-8410), human brain snRNA-seq data at GEO (GSE209552), human skin aging scRNA-seq data at GEO (GSE130973), PBMCs in SARS-CoV-2 infection scRNA-seq data at ArrayExpress (E-MTAB-9652).

Supplementary data

Supplementary Data are available at NAR Online.

Funding

Ministero della Salute [GR-2018-12365280 to B.B.]; Fondazione AIRC per la ricerca sul cancro ETS [IG 2022 27066 to B.B.]; Fondazione Cariplo [2019-3416 to B.B.]; Fondazione Regionale per la Ricerca Biomedica [CP2_12/2018 to B.B.]; Ministero dell'Università e della Ricerca [PNRR PE00000007 to B.B.]; Ministero dell'Università e della Ricerca [2022PKF9S to B.B.]; Fondazione Cariplo [2019-1788 to V.R.]. Funding for open access charge: Fondazione AIRC per la ricerca sul cancro ETS [IG 2022 27066 to B.B.], Progetti di Rilevante Interesse Nazionale (PRIN) [2022PKF9S to B.B.].

Conflict of interest statement

F.M. and B.B. are co-founders of the startup T-One Therapeutics s.r.l. All the other authors declare they have no competing interests.

References

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. 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.* (2018) Ten things you should know about transposable elements. *Genome Biol.*, **19**, 199.
3. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.
4. Sotero-Caio,C.G., Platt,R.N. II, Suh,A. and Ray,D.A. (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol. Evol.*, **9**, 161–177.
5. Lanciano,S. and Cristofari,G. (2020) Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, **21**, 721–736.
6. Chuong,E.B., Elde,N.C. and Feschotte,C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.
7. Attig,J., Agostini,F., Gooding,C., Chakrabarti,A.M., Singh,A., Haberman,N., Zagalak,J.A., Emmett,W., Smith,C.W.J., Luscombe,N.M., *et al.* (2018) Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. *Cell*, **174**, 1067–1081.
8. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T., *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.
9. Marasca,F., Sinha,S., Vadalà,R., Polimeni,B., Ranzani,V., Paraboschi,E.M., Burattin,F.V., Ghilotti,M., Crosti,M., Negri,M.L., *et al.* (2022) LINE1 are spliced in non-canonical transcript variants to regulate T cell quiescence and exhaustion. *Nat. Genet.*, **54**, 180–193.
10. Percharde,M., Lin,C.-J., Yin,Y., Guan,J., Peixoto,G.A., Bulut-Karslioglu,A., Biechele,S., Huang,B., Shen,X. and Ramalho-Santos,M. (2018) A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell*, **174**, 391–405.
11. Treangen,T.J. and Salzberg,S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
12. Sexton,C.E. and Han,M.V. (2019) Paired-end mappability of transposable elements in the human genome. *Mob. DNA*, **10**, 29.
13. Marasca,F., Gasparotto,E., Polimeni,B., Vadalà,R., Ranzani,V. and Bodega,B. (2020) The sophisticated transcriptional response governed by transposable elements in human health and disease. *Int. J. Mol. Sci.*, **21**, 3201.

14. Goerner-Potvin,P. and Bourque,G. (2018) Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.
15. Jin,Y., Tam,O.H., Paniagua,E. and Hammell,M. (2015) TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinforma.*, **31**, 3593–3599.
16. Yang,W.R., Ardeljan,D., Pacyna,C.N., Payer,L.M. and Burns,K.H. (2019) SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res.*, **47**, e27.
17. Kong,Y., Rose,C.M., Cass,A.A., Williams,A.G., Darwish,M., Lianoglou,S., Haverty,P.M., Tong,A.-J., Blanchette,C., Albert,M.L., *et al.* (2019) Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.*, **10**, 5228.
18. Bendall,M.L., Mulder,M.d., Iñiguez,L.P., Lecanda-Sánchez,A., Pérez-Losada,M., Ostrowski,M.A., Jones,R.B., Mulder,L.C.F., Reyes-Terán,G., Crandall,K.A., *et al.* (2019) Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.*, **15**, e1006453.
19. He,J., Babarinde,I.A., Sun,L., Xu,S., Chen,R., Shi,J., Wei,Y., Li,Y., Ma,G., Zhuang,Q., *et al.* (2021) Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nat. Commun.*, **12**, 1456.
20. Rodríguez-Quiroz,R. and Valdebenito-Maturana,B. (2022) SoloTE for improved analysis of transposable elements in single-cell RNA-Seq data using locus-specific expression. *Commun. Biol.*, **5**, 1063.
21. Stow,E.C., Baddoo,M., LaRosa,A.J., LaCoste,D., Deininger,P. and Belancio,V. (2022) SCIFER: approach for analysis of LINE-1 mRNA expression in single cells at a single locus resolution. *Mob. DNA*, **13**, 21.
22. Svensson,V., da Veiga Beltrame,E. and Pachter,L. (2020) A curated database reveals trends in single-cell transcriptomics. *Database*, **2020**, baaa073.
23. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
24. McKerrow,W., Kagermazova,L., Doudican,N., Frazzette,N., Kaparos,E.I., Evans,S.A., Rocha,A., Sedivy,J.M., Neretti,N., Carucci,J., *et al.* (2023) LINE-1 retrotransposon expression in cancerous, epithelial and neuronal cells revealed by 5′ single-cell RNA-Seq. *Nucleic Acids Res.*, **51**, 2033–2045.
25. Bonté,P.-E., Metoikidou,C., Heurtebise-Chretien,S., Arribas,Y.A., Sutra Del Galy,A., Ye,M., Niborski,L.L., Zueva,E., Piaggio,E., Seguin-Givelet,A., *et al.* (2023) Selective control of transposable element expression during T cell exhaustion and anti–PD-1 treatment. *Sci. Immunol.*, **8**, eadf8838.
26. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
27. Zheng,G.X.Y., 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.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
28. Kaminow,B., Yunusov,D. and Dobin,A. (2021) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. biorXiv doi: https://doi.org/10.1101/2021.05.05.442755, 05 May 2021, preprint: not peer reviewed.
29. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
30. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
31. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
32. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
33. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
34. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
35. Bonfield,J.K., Marshall,J., Danecek,P., Li,H., Ohan,V., Whitwham,A., Keane,T. and Davies,R.M. (2021) HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, **10**, giab007.
36. Hagberg,A.A., Schult,D.A. and Swart,P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. In: *Proc. 7th Python Sci. Conf. SciPy 2008*.
37. Frankish,A., Diekhans,M., Jungreis,I., Lagarde,J., Loveland,J.E., Mudge,J.M., Sisu,C., Wright,J.C., Armstrong,J., Barnes,I., *et al.* (2021) GENCODE 2021. *Nucleic. Acids. Res.*, **49**, D916–D923.
38. Lun,A.T.L., Riesenfeld,S., Andrews,T., Dao,T.P., Gomes,T., Marioni,J.C. and participants in the 1st Human Cell Atlas Jamboreeparticipants in the 1st Human Cell Atlas Jamboree (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, **20**, 63.
39. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 16 March 2013, preprint: not peer reviewed.
40. Hao,Y., Stuart,T., Kowalski,M.H., Choudhary,S., Hoffman,P., Hartman,A., Srivastava,A., Molla,G., Madad,S., Fernandez-Granda,C., *et al.* (2023) Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.*, **42**, 293–304.
41. Büttner,M., Miao,Z., Wolf,F.A., Teichmann,S.A. and Theis,F.J. (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, **16**, 43–49.
42. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
43. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
44. Garza,R., Atacho,D.A.M., Adami,A., Gerdes,P., Vinod,M., Hsieh,P., Karlsson,O., Horvath,V., Johansson,P.A., Pandiloski,N., *et al.* (2023) LINE-1 retrotransposons drive human neuronal transcriptome complexity and functional diversification. *Sci. Adv.*, **9**, eadh9543.
45. Solé-Boldo,L., Raddatz,G., Schütz,S., Mallm,J.-P., Rippe,K., Lonsdorf,A.S., Rodríguez-Paredes,M. and Lyko,F. (2020) Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun. Biol.*, **3**, 188.
46. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
47. van de Lagemaat,L.N., Landry,J.-R., Mager,D.L. and Medstrand,P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.*, **19**, 530–536.
48. Shao,W. and Wang,T. (2021) Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. *Genome Res.*, **31**, 88–100.
49. Zhu,X., Fang,H., Gladysz,K., Barbour,J.A. and Wong,J.W.H. (2021) Overexpression of transposable elements is associated with

immune evasion and poor outcome in colorectal cancer. *Eur. J. Cancer*, **157**, 94–107.

50. Jang,H.S., Shah,N.M., Du,A.Y., Dailey,Z.Z., Pehrsson,E.C., Godoy,P.M., Zhang,D., Li,D., Xing,X., Kim,S., *et al.* (2019) Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.*, **51**, 611–617.
51. Lee,H.-O., Hong,Y., Etlioglu,H.E., Cho,Y.B., Pomella,V., Van den Bosch,B., Vanhecke,J., Verbandt,S., Hong,H., Min,J.-W., *et al.* (2020) Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.*, **52**, 594–603.
52. Burns,K.H. (2017) Transposable elements in cancer. *Nat. Rev. Cancer*, **17**, 415–424.
53. Anwar,S.L., Wulaningsih,W. and Lehmann,U. (2017) Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *Int. J. Mol. Sci.*, **18**, 974.
54. Coufal,N.G., Garcia-Perez,J.L., Peng,G.E., Yeo,G.W., Mu,Y., Lovci,M.T., Morell,M., O'Shea,K.S., Moran,J.V. and Gage,F.H. (2009) L1 retrotransposition in human neural progenitor cells. *Nature*, **460**, 1127–1131.
55. Evrony,G.D., Lee,E., Mehta,B.K., Benjamini,Y., Johnson,R.M., Cai,X., Yang,L., Haseley,P., Lehmann,H.S., Park,P.J., *et al.* (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, **85**, 49–59.
56. Sanchez-Luque,F.J., Kempen,M.-J.H.C., Gerdes,P., Vargas-Landin,D.B., Richardson,S.R., Troskie,R.-L., Jesuadian,J.S., Cheetham,S.W., Carreira,P.E., Salvador-Palomeque,C., *et al.* (2019) LINE-1 Evasion of Epigenetic Repression in Humans. *Mol. Cell*, **75**, 590–604.
57. Yushkova,E. and Moskalev,A. (2023) Transposable elements and their role in aging. *Ageing Res. Rev.*, **86**, 101881.
58. Simon,M., Van Meter,M., Ablaeva,J., Ke,Z., Gonzalez,R.S., Taguchi,T., De Cecco,M., Leonova,K.I., Kogan,V., Helfand,S.L., *et al.* (2019) LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation. *Cell Metab.*, **29**, 871–885.
59. De Cecco,M., Ito,T., Petrashen,A.P., Elias,A.E., Skvir,N.J., Criscione,S.W., Caligiana,A., Brocculi,G., Adney,E.M., Boeke,J.D., *et al.* (2019) L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, **566**, 73–78.
60. Della Valle,F., Reddy,P., Yamamoto,M., Liu,P., Saera-Vila,A., Bensaddek,D., Zhang,H., Prieto Martinez,J., Abassi,L., Celii,M., *et al.* (2022) LINE-1 RNA causes heterochromatin erosion and is a target for amelioration of senescent phenotypes in progeroid syndromes. *Sci. Transl. Med.*, **14**, eabl6057.
61. Wang,J., Geesman,G.J., Hostikka,S.L., Atallah,M., Blackwell,B., Lee,E., Cook,P.J., Pasaniuc,B., Shariat,G., Halperin,E., *et al.* (2011) Inhibition of activated pericentromeric SINE/Alu repeat transcription in senescent human adult stem cells reinstates self-renewal. *Cell Cycle*, **10**, 3016–3030.
62. Tigges,J., Krutmann,J., Fritsche,E., Haendeler,J., Schaal,H., Fischer,J.W., Kalfalah,F., Reinke,H., Reifenberger,G., Stühler,K., *et al.* (2014) The hallmarks of fibroblast ageing. *Mech. Ageing Dev.*, **138**, 26–44.
63. Senapati,P., Miyano,M., Sayaman,R.W., Basam,M., Leung,A., LaBarge,M.A. and Schones,D.E. (2023) Loss of epigenetic suppression of retrotransposons with oncogenic potential in aging mammary luminal epithelial cells. *Genome Res.*, **33**, 1229–1241.
64. Bennett,E.A., Keller,H., Mills,R.E., Schmidt,S., Moran,J.V., Weichenrieder,O. and Devine,S.E. (2008) Active Alu retrotransposons in the human genome. *Genome Res.*, **18**, 1875–1883.
65. Kabanov,I.N., Mavropulo-Stolyarenko,G.R. and Tishchenko,L.I. (2018) Changes in gene expression and DNA methylation of evolutionarily young AluY repeats during apoptosis of human K562 erythro-myeloblastic leukemia cells. *J. Evol. Biochem. Physiol.*, **54**, 30–42.
66. Kitsou,K., Kotanidou,A., Paraskevis,D., Karamitros,T., Katzourakis,A., Tedder,R., Hurst,T., Sapounas,S., Kotsinas,A., Gorgoulis,V., *et al.* Upregulation of human endogenous retroviruses in bronchoalveolar lavage fluid of COVID-19 patients. *Microbiol. Spectr.*, **9**, e01260-21.
67. Marston,J.L., Greenig,M., Singh,M., Bendall,M.L., Duarte,R.R.R., Feschotte,C., Iñiguez,L.P. and Nixon,D.F. (2021) SARS-CoV-2 infection mediates differential expression of human endogenous retroviruses and long interspersed nuclear elements. *JCI Insight*, **6**, e147170.
68. Sorek,M., Meshorer,E. and Schlesinger,S. (2022) Impaired activation of transposable elements in SARS-CoV-2 infection. *EMBO Rep.*, **23**, e55101.
69. Yin,Y., Liu,X.-Z., Tian,Q., Fan,Y.-X., Ye,Z., Meng,T.-Q., Wei,G.-H., Xiong,C.-L., Li,H.-G., He,X., *et al.* (2022) Transcriptome and DNA methylome analysis of peripheral blood samples reveals incomplete restoration and transposable element activation after 3-months recovery of COVID-19. *Front. Cell Dev. Biol.*, **10**, 1001558.
70. Notarbartolo,S., Ranzani,V., Bandera,A., Gruarin,P., Bevilacqua,V., Putignano,A.R., Gobbini,A., Galeota,E., Manara,C., Bombaci,M., *et al.* (2021) Integrated longitudinal immunophenotypic, transcriptional and repertoire analyses delineate immune responses in COVID-19 patients. *Sci. Immunol.*, **6**, eabg5021.
71. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
72. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
73. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.*, **12**, 323.
74. McKerrow,W. and Fenyö,D. (2020) L1EM: A tool for accurate locus specific LINE-1 RNA quantification. *Bioinformatics*, **36**, 1167–1173.
75. Grillo,G., Keshavarzian,T., Linder,S., Arlidge,C., Mout,L., Nand,A., Teng,M., Qamra,A., Zhou,S., Kron,K.J., *et al.* (2023) Transposable elements are co-opted as oncogenic regulatory elements by lineage-specific transcription factors in prostate cancer. *Cancer Discov.*, **13**, 2470–2487.
76. Li,W., Prazak,L., Chatterjee,N., Grüninger,S., Krug,L., Theodorou,D. and Dubnau,J. (2013) Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat. Neurosci.*, **16**, 529–531.
77. Grillo,G. and Lupien,M. (2022) Cancer-associated chromatin variants uncover the oncogenic role of transposable elements. *Curr. Opin. Genet. Dev.*, **74**, 101911.
78. Ochoa Thomas,E., Zuniga,G., Sun,W. and Frost,B. (2020) Awakening the dark side: retrotransposon activation in neurodegenerative disorders. *Curr. Opin. Neurobiol.*, **61**, 65–72.

Received: December 20, 2022. **Revised:** August 22, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** September 2, 2024