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

Polimeni Benedetto[1,2], Marasca Federica[1,3], Ranzani Valeria[1*], Bodega Beatrice[1,4*]

[1]INGM, Istituto Nazionale di Genetica Molecolare 'Romeo ed Enrica Invernizzi', Milan, Italy.

[2]Ph.D. Program in Translational and Molecular Medicine, DIMET, University of Milan-Bicocca, Monza, Italy.

[3]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy.

[4]Department of Biosciences, University of Milan, Milan, Italy.

*Corresponding authors: bodega@ingm.org, ranzani@ingm.org

Abstract

Transposable elements (TEs) are interspersed and repetitive DNA sequences that contribute to the evolution of species by introducing novel regulatory sequences in the eukaryotic genome, and their transcription have been associated to both physiological functions and detrimental effects in human cells. Given their ubiquity and redundancy, the transcriptome analysis of TEs is notoriously challenging and requires specialized tools. While a plethora of software is available for bulk RNA-Seq, few attempts have been made for measuring the expression of TEs in scRNA-seq data so far. Here, we present IRescue (Interspersed Repeats single-cell quantifier), a software to quantify TE expression in scRNA-seq using a UMI deduplication algorithm to solve the allocation of reads ambiguously mapped on interspersed TEs. Using simulated data, we demonstrate that IRescue is precise at estimating the expression of TE subfamilies. Furthermore, by analysing normal and cancer cells from colorectal cancer (CRC) patients, we identify the TE expression signature of CRC cells, where

we found young LINE1 elements to be specifically enriched in subpopulations of cancer cells, compared to normal cells. Finally, we show the heterogeneity and expression dynamics of CRC TE markers previously characterized at bulk-level only, and that the expression of such markers is the result of the transcription of tumor-specific alternative isoforms of human oncogenes. With IRescue, we long to enable adding the layer of TE expression in the multimodal analysis of single cell sequencing and to facilitate the discovery of novel markers from the high and increasing amount of scRNA-seq experiments data.

Introduction

Transposable elements (TEs) are mobile genetic elements found in the genome of most eukaryotes. The genomic prevalence of TEs largely varies between organisms (1), and make approximately 46% and 32% of the human (2) and mouse (3) genomes, respectively. In most species, retrotransposons are by far the most abundant TEs, as they can mobilize and amplify themselves thanks to a “copy-and-paste” replication mechanism that results in a new copy of the retroelement in the host genome (4). Based on their insertion age, TEs have been hierarchically organized into classes (e.g. LINE, SINE, LTR), superfamilies (e.g. LINE1, Alu, ERVL) and families or subfamilies (e.g. L1PA2, AluY, HERVL) (5). The LINE1 superfamily comprehend the last few autonomously active and mobile TEs in the human genome (6, 7), while some Alu and SVA elements can mobilize by exploiting LINE1’s machinery (8). Besides these few active retrotransposons, the vast majority have lost the ability to generate new insertions; however, they can still be transcribed within other transcriptional units (host genes) and provide regulatory elements affecting gene expression and RNA processing (9). For example, evolutionarily old LINE1 fragments are included in tissue-specific alternative exons, represent a source of RNA-protein binding sites (10), and have been recently found to regulate the cell state maintenance of human T cells (11).

Next Generation Sequencing (NGS) technologies were indispensable to discover and annotate TE insertion in reference genomes and realizing the impact of TEs in genome and transcriptional regulatory networks. Yet, the study of TEs with high throughput sequencing is hampered by their extremely repetitive and interspersed nature (12). Some precautions during library design, such as choosing a paired-end layout or increasing the read length (13), and using software specifically developed to analyse interspersed repeats (14, 15), can drastically improve both the read mappability and expression estimate of TEs. Single cell sequencing technologies revolutionized the transcriptome analysis of complex tissues and organs, allowing the definition of the transcriptional profile of individual cells and the dissection of the tissue heterogeneity under different conditions. Unfortunately, single cell RNA sequencing (scRNA-seq) libraries, particularly those derived from droplet-based technologies (e.g. most Chromium 10x, Drop-seq and inDrops kits), are usually characterized by short reads with a strong 3'- or 5'-end positional bias. Moreover, these reads are effectively single-end, since the cDNA insert is only represented by one mate, while the other carries the cell barcode and unique molecule identifier (UMI) sequences. This tag-based layout decreases read mappability and makes it difficult to determine the exact genomic origin of TE-containing RNA fragments. Most of these limitations are overcomed by libraries prepared according to plate-based technologies, whose availability is steadily growing. However, droplet-based datasets are still by far predominant in public repositories (16); therefore, most of the publicly available datasets are difficult to be re-utilized for the analysis of TE expression without using specialized software.

Indeed, TE expression dynamics at single cell level is still poorly investigated, presumably due to the aforementioned computational limits and lack of dedicated software compatible with common scRNA-seq library layouts. Some of the recent efforts for the TE analysis at single cell level involved adapting existing tools for the quantification of TE expression (17), coupling bulk and single cell RNA-seq to reconstruct TE-containing transcripts (18) or developing ad-

hoc library preparation protocols (19). As of today, the only published software for the analysis of TEs at single cell level does not implement a UMI error correction and discards the information of the different TE subfamilies mapped by reads aligning multiple times on the genome, since it considers only one alignment per read (20). To rescue ambiguously mapping reads and UMIs and obtain better gene expression estimates, probabilistic-based approaches that leverage on building UMI-transcript equivalence classes to deduplicate and count UMIs have been developed (21, 22). These advancements achieved in bulk RNA-seq were recently carried over to single cell transcriptomics (23, 24); however, software using similar approaches to quantify TE expression in scRNA-seq is currently lacking. Following these principles, we developed IRescue (interspersed repeats single-cell quantifier), a command-line tool for the error-correction, deduplication and quantification of UMIs mapping on TEs in scRNA-seq using a UMI-TE equivalence class-based algorithm. IRescue is currently the only software that, in case of UMIs mapping multiple times on different TE subfamilies, takes into account all mapped features to estimate the correct one, rather than excluding multi-mapping UMIs or picking one randomly. In this study, we demonstrate the precision of IRescue using simulated data, identify TE expression signatures associated with colorectal cancer (CRC) and show the heterogeneity and expression dynamics of CRC TE markers previously characterized only at bulk-level. Moreover, we suggest that the expression of such markers is explained by the transcription of tumor-specific alternative isoforms of human oncogenes.

Materials and Methods

IRescue workflow

Input data. IRescue requires as input a BAM file containing read alignments on a reference genome, which must have the cell barcode and UMI sequences annotated as key-value tags. The keys for both the cell barcode and the UMI are by default the strings “CB” and “UR”, with the possibility to be overridden by the user. Additionally, it is mandatory to provide either the

name of a genome assembly (e.g. “hg38” for the human genome) or a BED file containing genomic TE coordinates. If a genome assembly name is provided, IRescue will automatically download and parse the Repeatmasker coordinates from the UCSC servers at the beginning of the workflow, filtering out unwanted short tandem repeats and repetitive RNA classes (i.e. Low_complexity, Simple_repeat, rRNA, scRNA, srpRNA, tRNA). If a custom BED file is provided, it is advised to use the TE subfamily as the feature name in the fourth column. The BED file must at least have four columns, can be plain text or compressed and the chromosome names must follow the same nomenclature as the genome assembly used for alignment. Repeatmasker annotations automatically downloaded by IRescue follow the UCSC nomenclature, i.e. with reference names having the “chr” prefix. Optionally, a cell barcode whitelist file (either plain text or compressed) can be provided in order to filter the reads with valid cell barcodes only. We suggest mapping reads with a spliced aligner that annotates alignments with the sequence-corrected cell barcodes, such as the “CB” tag added by CellRanger (25) or STARsolo (26), and to provide the filtered barcode list (i.e. the “barcodes.tsv” file) as a whitelist to IRescue, in order to obtain a TE count matrix consistent with the gene count matrix provided by these tools (supplementary figure 1b).

Mapping alignments on TEs. Read alignments and TE coordinates are processed in parallel by chromosome, according to the number of allocated CPUs. The intersection between read and TE coordinates is performed by IRescue wrapping bedtools 2.30.0 (27), taking into account eventual read deletions and splitting events due to splice junctions (i.e. TE loci localized between donor and acceptor splicing coordinates are not considered mapped). In case of reads associated with UMIs aligned on multiple TE loci, all the aligned TE features will be parsed and carried over for the UMI deduplication.

UMI deduplication algorithm. After mapping, cell barcodes are evenly splitted up in batches for parallel processing of single cell TE counts, according to the number of allocated CPUs. For each cell, UMIs and TE features are indexed and stored in memory as key-value pairs.

UMIs mapped on the same set of TE features (i.e. keys having the same value) are stored in the same equivalence class (EC), with the assumption that UMIs mapping on the same set of interspersed repetitive features carry an “equivalent” information regarding the mRNA molecule of origin. For each EC, the sequences of UMIs are compared in a pairwise manner to calculate the number of mismatches between them. Next, UMIs are arranged in an undirected graph where each node is a UMI and edges connect UMIs that differ for just one mismatch, as they are considered potential duplicates. To calculate the deduplicated number of UMIs explaining the EC, the algorithm finds the neighborhood of each node, i.e. a set containing the node itself and its adjacent nodes. We discern three basic possibilities of UMI networks. i) If a node has no adjacent nodes, thus a neighborhood of size (i.e. cardinality) one, it is considered a non-duplicated UMI, and adds 1 to the final UMI count of the EC. ii) If two or more nodes adjacent to each other all have the same neighborhood, they are considered derived from the same UMI, and adds 1 to the final UMI count of the EC. iii) In case of ambiguous connections, e.g. when three or more nodes that aren’t all adjacent are connected through a path, the algorithm finds the hubs of the graph, i.e. the nodes whose neighborhood’s intersection with any other neighborhood always results in a set of size smaller than the hub’s neighborhood’s size (supplementary figure 1c). Formally, given a set of n nodes, with $n \geq 3$, a neighborhood N_1 is a hub if the following comparison is always true:

$$|N_1 \cap N_n| < |N_1|$$

For example, in the graph in figure 1, the nodes 1 to 4 are connected on the same path, with nodes 1 and 3 being the network’s hubs, and adds a UMI count of 2, whereas the nodes 5 and 6 are a couple of adjacent nodes with the same neighborhood, hence derived from the same UMI (figure 1).

TE features count. Once the EC’s UMI count has been obtained, it is assigned to the TE feature showing the highest number of alignment events in the EC. Finally, the count of TE

features obtained from every EC of a cell is summed up, to obtain the final TE counts of the cell:

$$TEcount = \sum_{EC} TEcount_{EC}$$

The counts of TE features across cells are written in a TE×Cell sparse matrix following the Market Exchange format (MEX), which is compliant with the gene matrix output of Cell Ranger or STARsolo to ensure compatibility with several toolkits for single cell downstream analysis.

scRNA-seq alignment and gene quantification

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 (26, 28). As genes and splice junction database, we provided the Gencode comprehensive annotation v32 (29) with chromosome names converted to the UCSC nomenclature (passed to STAR through the "`--sjdbGTFfile`" parameter). We provided and the 10X Genomics cell barcodes whitelist for the v2 library kit (passed through the "`--soloCBwhitelist`" parameter). 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 --twopassMode Basic --soloType CB_UMI_Simple --soloCellFilter EmptyDrops_CR 10000 0.99 10 45000 90000 500 0.01 20000 0.01 10000`".

Single cell TE expression quantification

To estimate the expression of TEs at single cell level, read alignments were processed using IRescue 1.0.0, providing the filtered barcodes list produced by STARsolo as cell barcodes whitelist (passed through the "`--whitelist`" parameter), and other parameters "`--genome hg38 --CBtag CB --UMItag UR`". To compute TE counts with scTE 1.0 (20), 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`", in order to process the data consistently between the two tools.

Data simulations and benchmarks

Simulated scRNA-seq reads and TE subfamily counts were obtained adapting the method described by Kaminow et al. (26), which allows to reproduce the positioning bias of droplet-based scRNA-seq on genes, introns and intergenic regions by using a real dataset as the basis for the simulation. For this purpose, we used a 10X Genomics v2 3'-end PBMC dataset (https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0), which is a library layout common among publicly available scRNA-seq datasets. Briefly, cell barcodes are filtered according to the 10X whitelist and UMIs with uncalled bases are removed, reads are aligned on a reference that combined the human hg38 primary genome assembly and the Repeatmasker TE genomic sequences using BWA-MEM 0.7.17 (30). UMIs are counted based on the TE subfamily they map on; in case alignments on multiple features, the top-scoring is chosen. Finally, the reference sequence is extracted from the aligned coordinates and a mismatch error rate of 0.5% was added to simulate Illumina sequencing errors. Summarizing plots of multimapping UMIs and equivalence classes (figure 2a-c) were obtained from the UMI-TE mappings kept by IRescue as temporary files with the "`--keep-tmp`" flag. Spearman correlation tests were done using the R 4.1.0 stats package between IRescue and simulated TE counts or features per cell (figure 2d and 2e, respectively), or between IRescue and scTE total counts or counts by cell (supplementary figures 2a and 2c, respectively). The relative deviation (RD) between IRescue or scTE measured counts and simulated counts was calculated as:

$$RD = \frac{measured - simulated}{max(measured, simulated)}$$

and represented on a histogram with 0.001 bin size (supplementary figure 2b). Cell clustering according to TE counts was performed using Seurat 4.0.5 (31) and Louvain algorithm, and the comparison between IRescue or scTE and simulated clusters was performed with a chi squared test, and visualized by plotting the Pearson's residuals using the R corrplot 0.92 package (figure 2f, supplementary figure 2d). Memory (RAM) usage and run time were measured by running IRescue and scTE in a Nextflow pipeline (32) and extracted from the pipeline's report (supplementary figures 2e-f).

Analysis of scRNA-seq cancer dataset

Colorectal cancer and CRC-adjacent cells scRNA-seq raw data were obtained from ArrayExpress E-MTAB-8410 (33), along with the annotation of cell identity based on gene expression. The epithelial cells from six paired tumor and normal samples were analyzed using the Seurat 4.0.5 toolkit. Cells TE counts were normalized and scaled, then PCA and UMAP dimensionality reductions were performed and 12 clusters were discovered using the Louvain algorithm. UMAP representations of TE counts were plotted marking the cells according to their annotated identity (figure 3a) or TE expression cluster (supplementary figure 3a). TE expression clusters were identified as predominantly tumoral (K) or normal (N) by inspecting the abundance of cell identities on each cluster (figure 3b). TE expression signatures were obtained by finding differentially expressed TEs between each cluster and the rest of the dataset, and TEs significatively expressed (Wilcoxon rank sum test's adjusted p-value < 0.05; log2 fold change > 1) on K or N clusters only were selected (supplementary figures 3b-c). The difference between the abundance of TE classes in tumor and normal signatures was tested with a two-tailed two-proportions Z-test using the R 4.1.0 stat package (figure 3c). The average expression of CRC TE markers across conditions or clusters (figure 2e-f) was visualized using Seurat 4.0.5. Sashimi plots (figure 2g, supplementary figure 2e) were obtained by importing the alignments (BAM file) and TE coordinates on the IGV genome browser (34).

Results

An algorithm for TE expression quantification using ambiguously mapped UMIs in scRNA-seq

Rescuing multimapping reads is well known to be a compulsory procedure in order to accurately quantify the expression of TEs (35) or other poorly mappable features, such as multigene families (36). In this context, it has been shown that randomly selecting one alignment position of multimapping reads yields sufficiently precise expression estimates of TEs at subfamily level in bulk high-throughput sequencing (37), suggesting that the possibility of a wrong allocation of a read is mitigated by the high sequencing depth per sample achievable. However, using this strategy in droplet-based scRNA-seq could lead to false positives or negatives, due to the lower sequencing depth per cell, and the consequent limited amount of TE-associated UMIs per cell, compared to bulk RNA-seq. Hence, one of the cornerstones that motivated the development of IRescue was to use the information of all mapped features to allocate ambiguously mapped UMIs with higher confidence. Given that most expressed TEs in mammalian genomes are small fragments transcribed in association to other genes (38), many reads align on both gene's exons and TEs (supplementary figure 1a). Thus, to avoid arbitrarily allocating these reads to either genic or TE features, potentially missing out relevant biological signal, we designed IRescue to be coupled to a scRNA-seq gene expression quantifier of choice, to allow integrating gene and TE counts in downstream analysis (supplementary figure 1b). The quantitative analysis of RNA by quantifying UMIs is an established concept for detecting PCR or optical duplicates in high-throughput sequencing, and for this purpose, several tools can correct small sequencing errors in the sequence of UMIs, to avoid overestimating the UMI counts (25, 39). However, there are no methods available to correct sequencing errors in multimapping UMIs that align on more than one feature or meta-features, such as TE subfamilies. Hence, we implemented in IRescue a novel UMI deduplication algorithm that takes into account for mismatches between both uniquely

and multimapping UMIs. IRescue takes as input a binary aligned map (BAM) file (40), containing reads aligned on a reference genome assembly with cell barcode and UMI annotated as tags of type string, and a browser extensible data (BED) file (41) containing genomic TEs coordinates. While any spliced aligner generating an output with these features will be compatible with IRescue, as of today we recommend using STARsolo (26, 28) with custom parameters optimized for reads alignment on interspersed repetitive elements (see methods). IRescue is programmed to download and parse into a BED file the genomic Repeatmasker TE coordinates (42) of the genome assembly of choice directly from the UCSC server (41); alternatively, it is possible to use a custom BED file with TE coordinates. 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 (25) to ensure compatibility with most toolkits for downstream analysis (31, 39) (supplementary figure 1b). IRescue extracts all the alignments on TEs and parses the cell barcode and UMI sequences, as well as the mapped TE features; then, computes the TE counts for each cell by running an algorithm for the correction of multimapping reads and duplicated UMIs. Briefly, UMIs mapped within a cell on the same set of TE features are grouped into *n* equivalence classes (ECs); then, for each EC of a cell, duplicated UMIs are detected by inspecting their sequence similarity, admitting up to one mismatch in UMI sequence to find duplicates. After UMI deduplication, the final UMI count for an EC is assigned to the TE feature showing the highest amount of alignments (in case of ties, the count gets evenly splitted among the features). The counts of TE features from all ECs within a cell are summed up to obtain the cell's final TE counts. Finally, the TE counts of all the sample's cells are collected and written into a TE×Cell matrix (figure 1). Since the detection of invalid cell barcodes is a problem already addressed by several methods for gene expression quantification in scRNA-seq (21, 22, 25, 26, 44), and since the use case of IRescue is to integrate TE expression to common scRNA-seq gene expression analyses, we decided to not implement a cell barcode whitelisting strategy in IRescue; instead, it is possible to give as additional input to IRescue a list of valid cell barcodes. We strongly suggest providing the

cell barcode whitelist generated by the gene expression quantification method of choice, in order to have a library of cells consistent between gene and TE counts. In this paper, we provided as a whitelist the filtered barcodes list obtained from STARsolo during the alignment pre-processing step using the EmptyDrops algorithm (44). IRescue is written in the Python programming language and leverages on essential open source libraries for efficient scientific computing and bioinformatics data wrangling (27, 40, 45–47).

scRNA-seq simulations confirms that IRescue is accurate in quantifying TE expression at single-cell resolution

To evaluate the precision of IRescue's quantification strategy, we sought to generate simulated scRNA-seq reads carrying the same biases of real datasets, such as reads mapping to multiple genomic loci and to intronic and intergenic regions, including TE fragments spliced in unknown transcript isoforms or autonomously transcribed TEs. To achieve this, we implemented the method described by Kaminow et al. (26), in which they used a real dataset as a basis for the generation of simulated reads and counts, with custom modifications in order to obtain simulated counts at TE feature-level, rather than gene-level (see methods). As the basis for the simulation, we used a typical 10X Genomics v2 protocol library of peripheral blood mononuclear cells, containing 8,341 cells with ~92,000 reads per cell, 16bp long cell barcodes, 10bp UMIs and 98bp 3'-end cDNA sequences. After aligning simulated reads on the reference genome, we assessed the mappability over TEs at UMI-level and found that, of the 33% UMIs mapped on TEs, only 21% were uniquely mapped, while 39% mapped multiple times on the same TE subfamily and the remaining 40% on different TE subfamilies (figure 2a). Next, we obtained TE counts from the aligned simulated reads using IRescue. To evaluated the distribution and composition of ECs across cells, we processed IRescue's temporary files and observed that most cells contains ~1000 ECs, each consisting of a unique set of UMIs mapping on the same set of TE subfamilies (figure 2b), and that most ECs contains 10 or less different TE subfamilies (figure 2c). Therefore, UMIs mapping on more than 10

distinct TE subfamilies are rare. To test the accuracy of estimating TE expression in single cells, we inspected the correlation between simulated and measured TE counts (figure 2d) and number of TE features (figure 2e) and found it very high in both measurements. To compare IRescue's precision with other existing methods, we repeated the analysis on simulated data using scTE (20), and found a slightly stronger correlation between measured and simulated counts using IRescue, compared to scTE (supplementary figure 2a). Furthermore, the vast majority of TE counts measured across all cells showed a minimal deviation from the simulated counts in both IRescue and scTE (supplementary figure 2b), whereas per-cell TE counts were found to correlate better to simulated counts in a greater amount of cells when using IRescue, compared to scTE (supplementary figure 2c). To test the ability to infer the identity of single cells in respect to their TE expression profile, we tested the performance of dimensionality reduction and cell clustering techniques on measured and simulated data. We applied a canonical workflow of scRNA-seq downstream analysis using the Seurat toolkit (31, 48) in order to group cells according to their TE expression profile. We found that clusters of cells analysed with IRescue are equivalent to their simulated counterpart, with only a partial collision observed between clusters 0 and 2 (figure 2f). Performing the same analysis with scTE, we found that two of nine identified clusters completely failed to match a correspondent identity in simulated data, showing that IRescue performs better in identifying the cell's state in respect to TE expression profile (supplementary figure 2d). Finally, we analysed the resource usage of IRescue and scTE by the number of processing units (CPUs) utilized. While scTE had a shorter run time when using a low number of CPUs, IRescue's run times scaled better with an increasing number of CPUs (supplementary figure 2e); on the other hand, IRescue showed a very low memory footprint compared to scTE (supplementary figure 2f). The higher run time on IRescue was expected, as it performs more tasks than scTE (including UMI mismatch correction and weighting of multi-mapped TEs). Because IRescue is programmed to evenly split up the cells according to the allocated CPUs for parallelizing the count process, the run time drastically decreases when using a higher number of CPUs. The

low memory footprint of IRescue is achieved by an efficient use of Python data structures for the indexing of cell barcodes, UMIs and TE features, and by performing operations leveraging on Python generators with a first-in-first-out logic.

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

TEs are known to be dysregulated in cancer, where they can initiate oncogene expression (49) or cause genomic instability (50, 51). Several studies reported the overexpression of specific TE subfamilies in cancer cells with bulk-level NGS analyses (49, 52, 53); however, the expression dynamics of TEs in cancer at single cell level and their contribution to the heterogeneity of cell subsets is still poorly characterized. Therefore, to demonstrate IRescue performance in real datasets, we analysed publicly available 10X 3' scRNA-seq datasets from tumor and normal adjacent tissues of six colorectal cancer (CRC) patients (33). We performed an unsupervised clustering analysis on 32,276 tumor and normal epithelial cells based on TE counts computed by IRescue. Interestingly, we found that cell clustering based on TE expression alone discerns cancer from normal cells (figure 3a, supplementary figure 3a). In particular, we verified that, of 12 inferred clusters, half are populated exclusively or mostly of cancer cells (K clusters) and the other half of normal cells (N clusters, figure 3b), showing that TE expression profiles were indicative of most cells' condition. To identify the TEs responsible for such behaviour, we obtained the TE expression signatures of tumor and normal cells clusters by performing a differential expression analysis and selecting the TEs overexpressed only in one of the two conditions. We found the expression of TE subfamilies to be more consistent in cancer clusters compared to the normal ones (supplementary figure 2b-c). Interestingly, we found that LINE is the only class significantly over-represented in the TE expression signature of cancer cells (figure 3c). By inspecting the differentially expressed LINE subfamilies, we observed that they are all LINE1 elements; moreover, the most evolutionarily young ones (i.e. L1HS and L1PA*) are overexpressed in cancer cells only (figure 3d), hinting

to a possible de-repression of such young insertions in cancer conditions. As a validation to our differential analysis, we checked the presence of 16 TE features known to be overexpressed in CRC from bulk RNA-seq studies (52, 53), and confirmed the presence of 10 out of 16 CRC TE markers among the TE subfamilies belonging to our cancer TE signature (figure 3e). Furthermore, thanks to the single-cell resolution, we were able to reveal a previously uncharacterized heterogeneous distribution of CRC TE markers (figure 3f). For instance, L1PA2 and HERVH-int, both known to be associated with CRC (53), are specific for different K clusters (supplementary figure 3d). Next, we asked whether it was possible to infer, under certain conditions, the origin of the TE overexpression in cancer cells clusters. Since a specific L1PA2 insertion was reported to be spliced into an alternative isoform the oncogene SYT1 in CRC (49), we inspected the read alignments over this particular locus, specifically looking for reads splitted across a splice junction mapped on both TEs and non-repetitive genomic regions that could support the existence of cryptic TE-containing exons. Indeed, reconstructing the entire structure of unannotated transcripts is unfeasible in droplet-based scRNA-seq but it is still possible to inspect the usage of splice junctions (26). We found reads mapping on L1PA2 matching the coordinates of a previously reported cryptic exon (49) with higher coverage in tumor cells, compared to normal ones. In particular, we found over 3-fold more reads splitted between the L1PA2 exon and the 3rd exon of SYT1 in tumor (figure 3g), supporting the detection of the same previously reported L1PA2-derived tumor-specific cryptic exon (49). Likewise, we confirm the detection of a MER1B-PIWIL1 TE-oncogene isoform reported in CRC (49) (supplementary figure 3e). These results described the complex dynamics of TE expression in CRC, that involve specific TE subfamilies across groups of cells, with a resolution that was not possible to achieve by previous works based on bulk-level NGS strategies. Moreover, these findings support the notion that the differential expression of TE subfamilies in cancer cells results from tumor-specific alternative splicing events of TE-containing transcripts.

Discussion

In this work, we introduced IRescue, the first software for error-correction, deduplication and quantification of UMIs mapping on multiple interspersed genomic features. It is known that reads originating from repetitive elements align equally well on multiple genomic loci (35), and aligners usually report either one random alignment per read, or one arbitrarily designated primary alignment and a number of secondary alignments. As a consequence, discarding secondary alignments could lead to an incorrect feature assignment for that read. Several tools for TE expression quantification in bulk RNA-seq have been developed to take into account secondary alignments when assigning multi-mapping reads to features (14, 35). Such strategies have not yet been implemented in scRNA-seq, where only few methods for gene expression quantification are able to recover multi-mapping reads (21, 22, 26). In this context, IRescue not only takes into account all the secondary alignments of multi-mapping reads for feature assignment, but also provides the first strategy for the deduplication of UMIs multi-mapping on a reference genome. As a solution to assign a unique feature to multi-mapping UMIs, we leveraged on equivalence classes (ECs), a well-known concept used to count reads compatible to multiple isoforms in bulk RNA-seq (23, 24, 54), recently applied to gene expression quantification in scRNA-seq as well (21, 22). We generated simulated 10X reads and TE feature counts and showed that, despite the high amount of ambiguous alignments, IRescue is highly precise when quantifying TE expression at subfamily level. We showed that, in a common 10X scRNA-seq library layout, one third of UMIs are TE-associated, of which 40% aligns on multiple TE subfamilies. This observation indicates that alignments on multiple TE meta-features are frequent and should be considered in TE expression quantification, even when performing a count at subfamily level. We compared IRescue to scTE (20), finding that, while TE counts are overall similar to simulations for both tools, IRescue performs better, particularly when inferring the cell's identity using a common scRNA-seq clustering algorithm. We deem that this result is a consequence of scTE discarding the information of UMIs mapping to multiple TE subfamilies, an event that we showed being frequent in droplet-based scRNA-

seq. In the context of TE overexpression in cancer, we analysed the single cell TE expression profiles of CRC and the tumor-adjacent tissue, and found an enrichment in evolutionarily young LINE subfamilies in the CRC TE signature. Furthermore, we showed that known CRC TE markers, previously characterized only at bulk level RNA-seq, are heterogeneously expressed in different CRC cell clusters, and that the expression of such markers can be the result of the transcription of tumor-specific alternative isoforms of human oncogenes (49). The transcription of RNA containing young LINE insertions in healthy somatic cells have been shown to be preferentially repressed during RNA processing, in contrast to old insertions that have been slowly co-opted by species throughout evolution (10). The transcription of TEs has been shown to be involved in several physiological scenarios, such as embryo development (55), aging (56), differentiation and cell state maintenance (11), as well as pathological conditions, such as cancer (57) and neurodegenerative diseases (58); 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 experiment, enabling the possibility of extracting novel information from the high amount of publicly available datasets (16).

Data Availability

IRescue source code and documentation are available at https://github.com/bodegalab/irescue. Data used in this paper are available at 10X Genomics (https://www.10xgenomics.com/resources/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0) and ArrayExpress (E-MTAB-8410).

References

1. Wells,J.N. and Feschotte,C. (2020) A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet*, **54**, 539–561.

2. 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.

3. Chinwalla,A.T., Cook,L.L., Delehaunty,K.D., Fewell,G.A., Fulton,L.A., Fulton,R.S., Graves,T.A., Hillier,L.W., Mardis,E.R., McPherson,J.D., *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

4. Kazazian,H.H. (2004) Mobile Elements: Drivers of Genome Evolution. *Science*, **303**, 1626–1632.

5. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

6. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*, **100**, 5280–5285.

7. Beck,C.R., Collier,P., Macfarlane,C., Malig,M., Kidd,J.M., Eichler,E.E., Badge,R.M. and Moran,J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, **141**, 1159–1170.

8. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet*, **35**, 41–48.

9. 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.

10. 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.e17.

11. 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.

12. Treangen,T.J. and Salzberg,S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, **13**, 36–46.

13. Sexton,C.E. and Han,M.V. (2019) Paired-end mappability of transposable elements in the human genome. *Mobile DNA*, **10**, 29.

14. 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. *International Journal of Molecular Sciences*, **21**, 3201.

15. Goerner-Potvin,P. and Bourque,G. (2018) Computational tools to unmask transposable elements. *Nat Rev Genet*, **19**, 688–704.

16. Svensson,V., da Veiga Beltrame,E. and Pachter,L. (2020) A curated database reveals trends in single-cell transcriptomics. *Database*, **2020**, baaa073.

17. Brocks,D., Chomsky,E., Mukamel,Z., Lifshitz,A. and Tanay,A. (2019) Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. 10.1101/462853.

18. 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.

19. Berrens,R.V., Yang,A., Laumer,C.E., Lun,A.T.L., Bieberich,F., Law,C.-T., Lan,G., Imaz,M., Bowness,J.S., Brockdorff,N., *et al.* (2022) Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nat Biotechnol*, **40**, 546–554.

20. 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.

21. Srivastava,A., Malik,L., Smith,T., Sudbery,I. and Patro,R. (2019) Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology*, **20**, 65.

22. Melsted,P., Booeshaghi,A.S., Liu,L., Gao,F., Lu,L., Min,K.H. (Joseph), da Veiga Beltrame,E., Hjörleifsson,K.E., Gehring,J. and Pachter,L. (2021) Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol*, **39**, 813–818.

23. 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.

24. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, **34**, 525–527.

25. 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.

26. Kaminow,B., Yunusov,D. and Dobin,A. (2021) STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. 10.1101/2021.05.05.442755.

27. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

28. 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.

29. 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.

30. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 10.48550/arXiv.1303.3997.

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.e29.

32. 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.

33. 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.

34. 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.

35. Lanciano,S. and Cristofari,G. (2020) Measuring and interpreting transposable element expression. *Nat Rev Genet*, **21**, 721–736.

36. 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.

37. Teissandier,A., Servant,N., Barillot,E. and Bourc'his,D. (2019) Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mobile DNA*, **10**, 52.

38. 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.

39. 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.

40. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

41. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler, and D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

42. Smit,A., Hubley,R. and Green,P. (2013) RepeatMasker Open-4.0.

43. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, **19**, 15.

44. Lun,A.T.L., Riesenfeld,S., Andrews,T., Dao,T.P., Gomes,T., Marioni,J.C., and participants in the 1st Human Cell Atlas Jamboree (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, **20**, 63.

45. 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.

46. 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.

47. Heger,A. and Jacobs,K. (2020) Pysam.

48. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*, **33**, 495–502.

49. 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.

50. Kines,K.J., Sokolowski,M., deHaro,D.L., Christian,C.M. and Belancio,V.P. (2014) Potential for genomic instability associated with retrotranspositionally-incompetent L1 loci. *Nucleic Acids Res*, **42**, 10488–10502.

51. Goodier,J.L. (2016) Restricting retrotransposons: a review. *Mobile DNA*, **7**, 16.

52. 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. *European Journal of Cancer*, **157**, 94–107.

53. 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.

54. Nicolae,M., Mangul,S., Măndoiu,I.I. and Zelikovsky,A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, **6**, 9.

55. 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.e19.

56. 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.

57. Burns,K.H. (2017) Transposable elements in cancer. *Nat Rev Cancer*, **17**, 415–424.

58. Ochoa Thomas,E., Zuniga,G., Sun,W. and Frost,B. (2020) Awakening the dark side: retrotransposon activation in neurodegenerative disorders. *Current Opinion in Neurobiology*, **61**, 65–72.

Figures

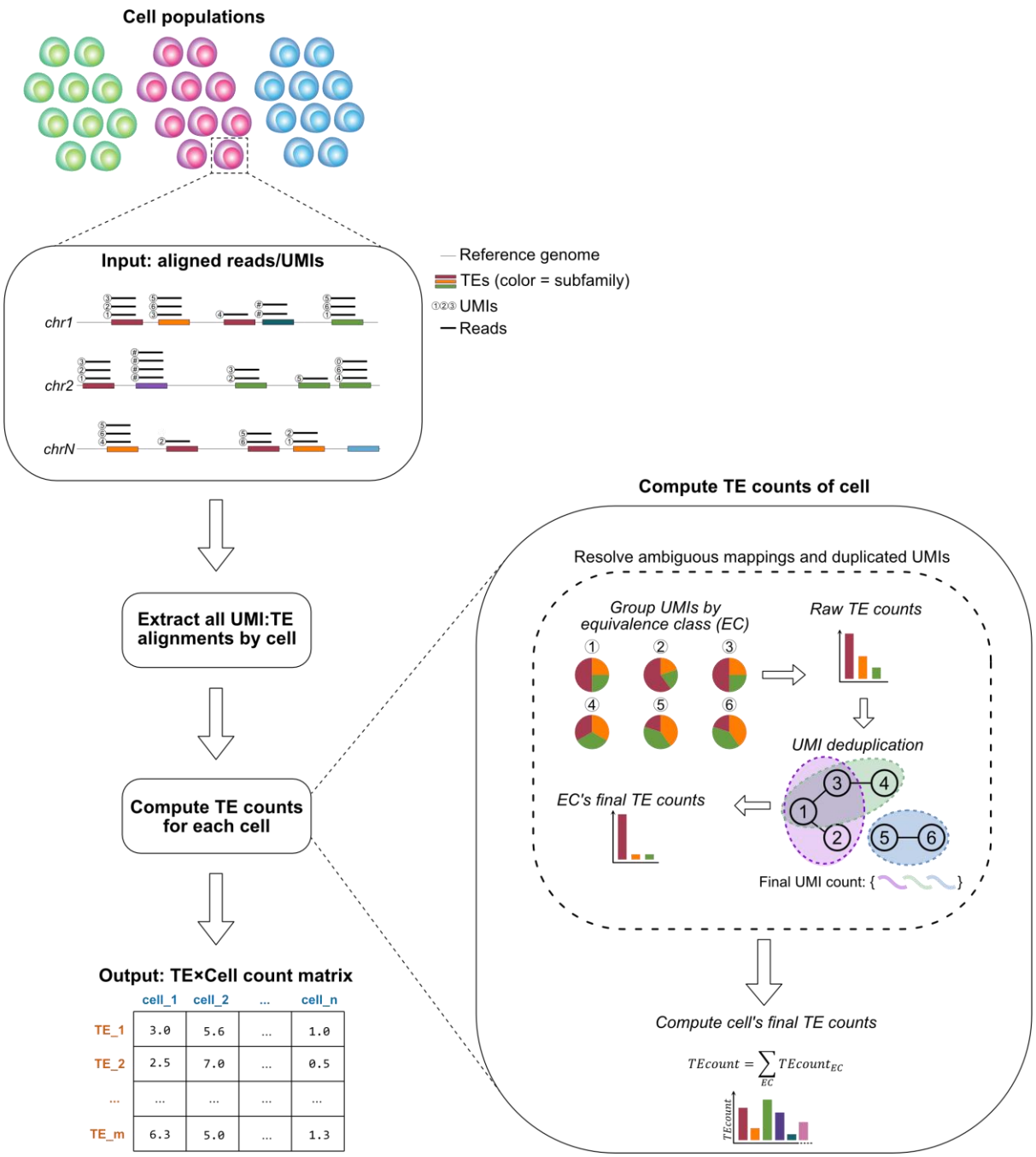


Figure 1. IRescue algorithm overview. scRNA-seq reads aligned on a reference genome, and the associated UMIs and cell barcodes, are parsed by IRescue and, for each cell, UMIs mapped on the same set of TEs are grouped into equivalence classes (ECs). For each EC, the initial UMI count of TEs is calculated considering all multiple alignments on different TE subfamilies. The actual final UMI count of the equivalence classes is inferred by deduplicating

the UMIs based on their sequence similarity. The deduplicated UMI count is assigned on the TE subfamily with the highest amount of alignments (in case of ties, the count is splitted between subfamilies). The counts of all equivalence classes of a cell are summed together to obtain the cell's final TE counts. The TE counts are written in a m×n matrix, with m and n being the number of TEs and cells, respectively.

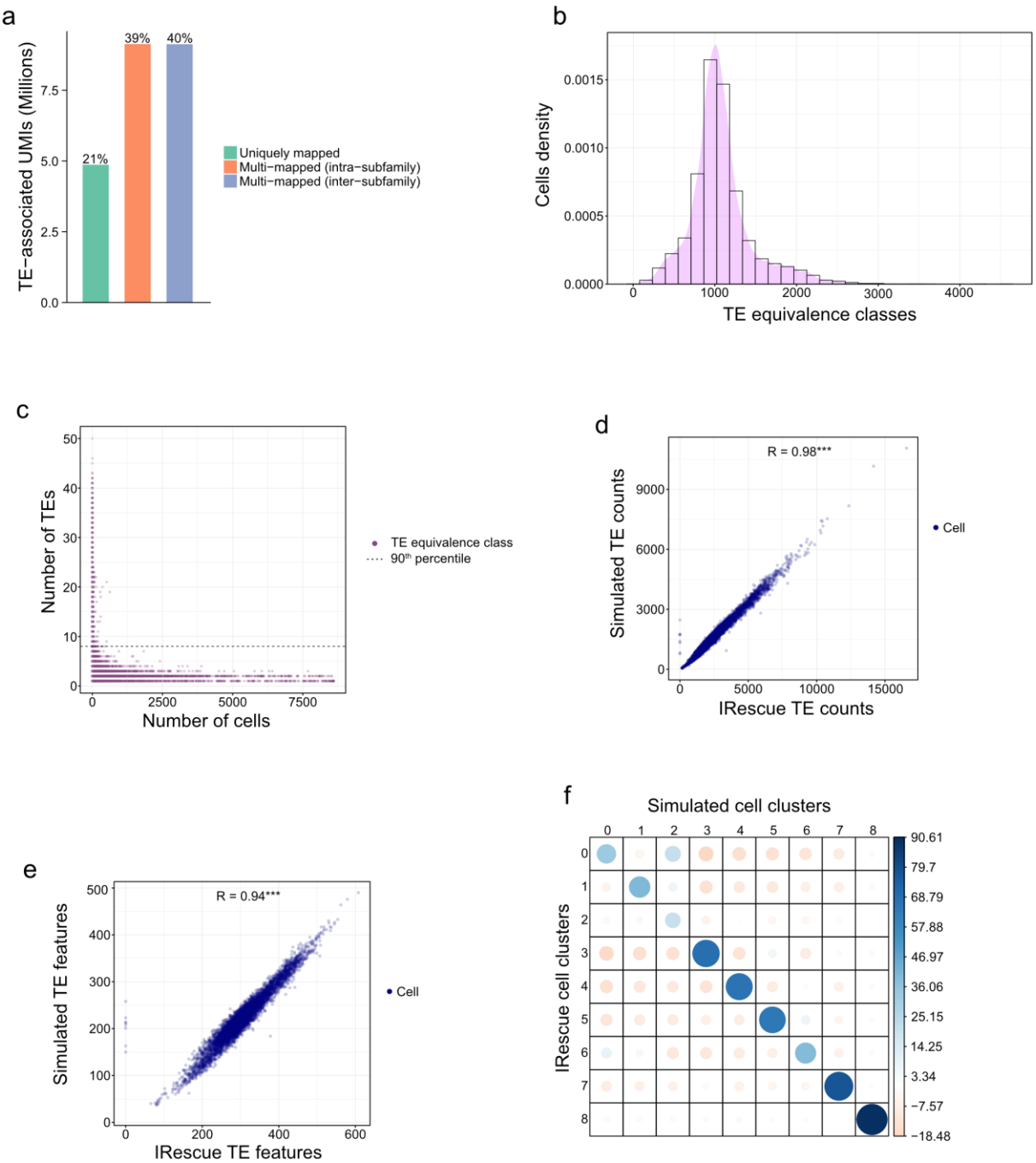


Figure 2. Performance evaluation of IRescue using simulated scRNA-seq data. a. Count of TE-associated UMIs, divided into uniquely mapped (i.e. mapped only once in the genome), multi-mapped across multiple loci of the same TE subfamily or across different TE subfamilies. **b.** Distribution of the count of TE equivalence classes detected across cells. **c.** Distribution of TE equivalence classes according to the number of cells on which they are detected and the number of TEs they contain (Dashed line = 90th percentile; R = Spearman correlation coefficient; *** = Spearman's rank correlation test's p-value < 2.2×10^{-16}). **d-e**. Relationship between IRescue- and simulated- TE counts (d) or TE features (e). *** as in (c). **f.** Correlation matrix plot between cell clusters inferred from IRescue and simulated data, using the Louvain algorithm implemented in Seurat v4. Pearson's correlation residuals of all pairwise combinations are plotted as dots shaded blue or red depending on the correlation being positive or negative. Color intensity and dot size are proportional to the magnitude of the correlation. p-value < 2.2×10^{-16} (chi-squared contingency table test).

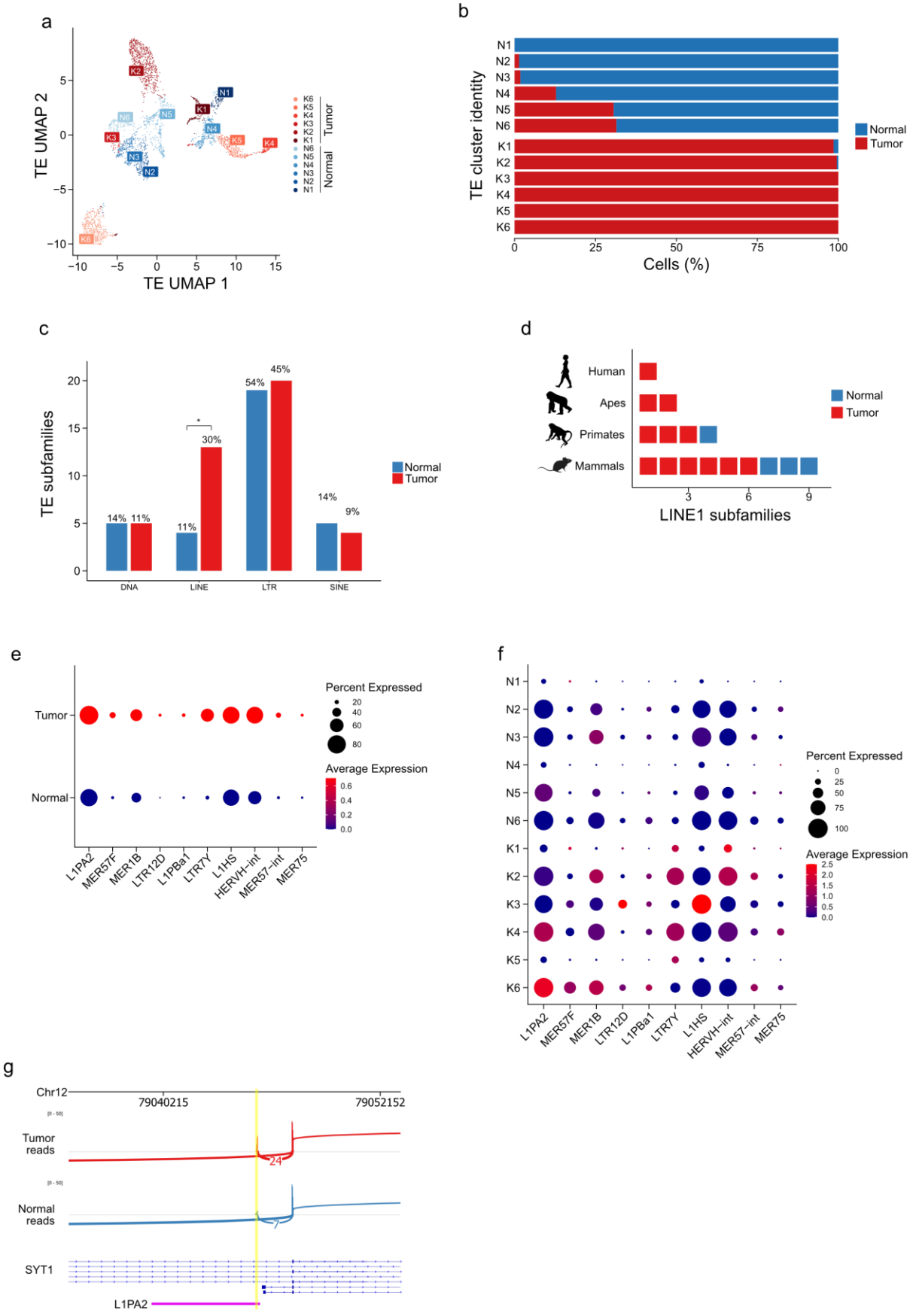
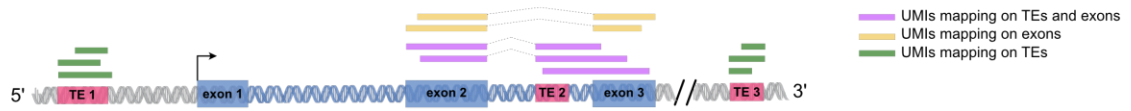
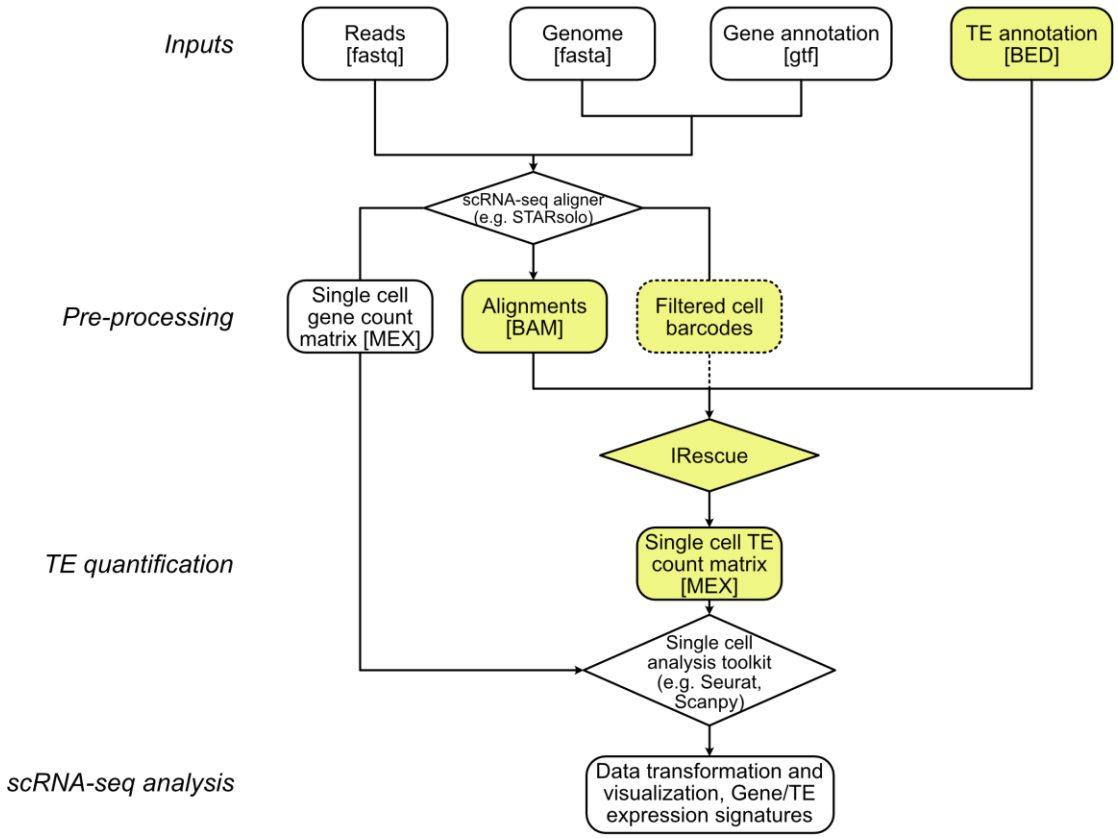

a
TE UMAP 2
TE UMAP 1
K6
K5
K4
K3
K2
K1
N6
N5
N4
N3
N2
N1
Tumor
Normal
b
TE cluster identity
Cells (%)
Normal
Tumor
c
TE subfamilies
DNA
LINE
LTR
SINE
14%
11%
11%
30%
54%
45%
14%
9%
Normal
Tumor
d
Human
Apes
Primates
Mammals
LINE1 subfamilies
Normal
Tumor
e
Tumor
Normal
L1PA2
MER57F
MER1B
LTR12D
L1PBa1
LTR7Y
L1HS
HERVH-int
MER57-int
MER75
Percent Expressed
Average Expression
f
N1
N2
N3
N4
N5
N6
K1
K2
K3
K4
K5
K6
L1PA2
MER57F
MER1B
LTR12D
L1PBa1
LTR7Y
L1HS
HERVH-int
MER57-int
MER75
Percent Expressed
Average Expression
g
Chr12
79040215
79052152
Tumor reads
Normal reads
SYT1
L1PA2

Figure 3. Identification of TE expression dynamics in colorectal cancer. a. UMAP representation of CRC and normal cells according to TE expression. Clusters of normal and tumour cells are obtained on the basis of TE expression. **b.** Barplot of the relative abundance of cells by condition across TE clusters. **c.** Barplot of abundance and enrichment of TE subfamilies by TE class in tumor and normal TE signatures. * p-value < 0.05 (two-sided two-proportions Z-test). **d**. Representation of LINE1 subfamilies specific for cancer or normal cells, stratified by the phylogenetic clade in which the insertions originated **e**. Dot plot of the average expression of known TE CRC markers in tumor and normal conditions. The dot size is indicative of the percentage of expressing cells. **f**. Dot plot of the average expression of known TE CRC markers across TE clusters. The dot size is indicative of the percentage of expressing cells. **g.** Sashimi plot representing the coverage of a L1PA2-derived CRC-specific cryptic exon of SYT1 oncogene and the splitted reads across the cryptic and canonical exon.

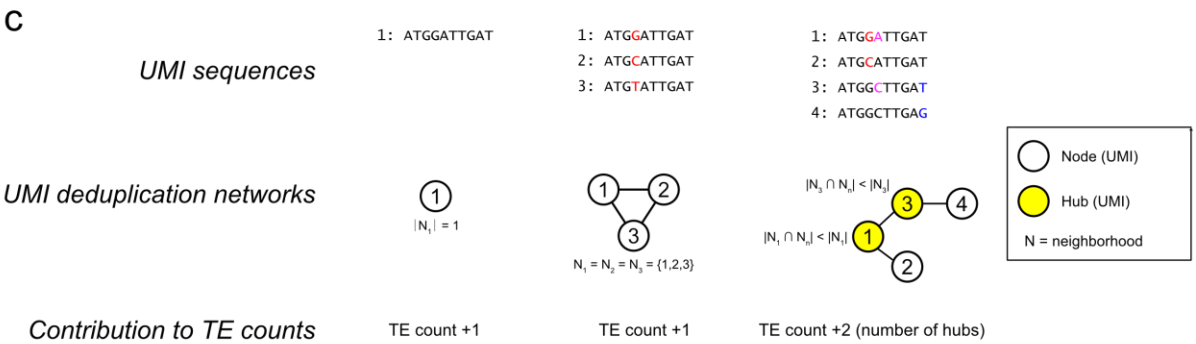
a


UMIs mapping on TEs and exons
UMIs mapping on exons
UMIs mapping on TEs
5'
TE 1
exon 1
exon 2
TE 2
exon 3
TE 3
3'

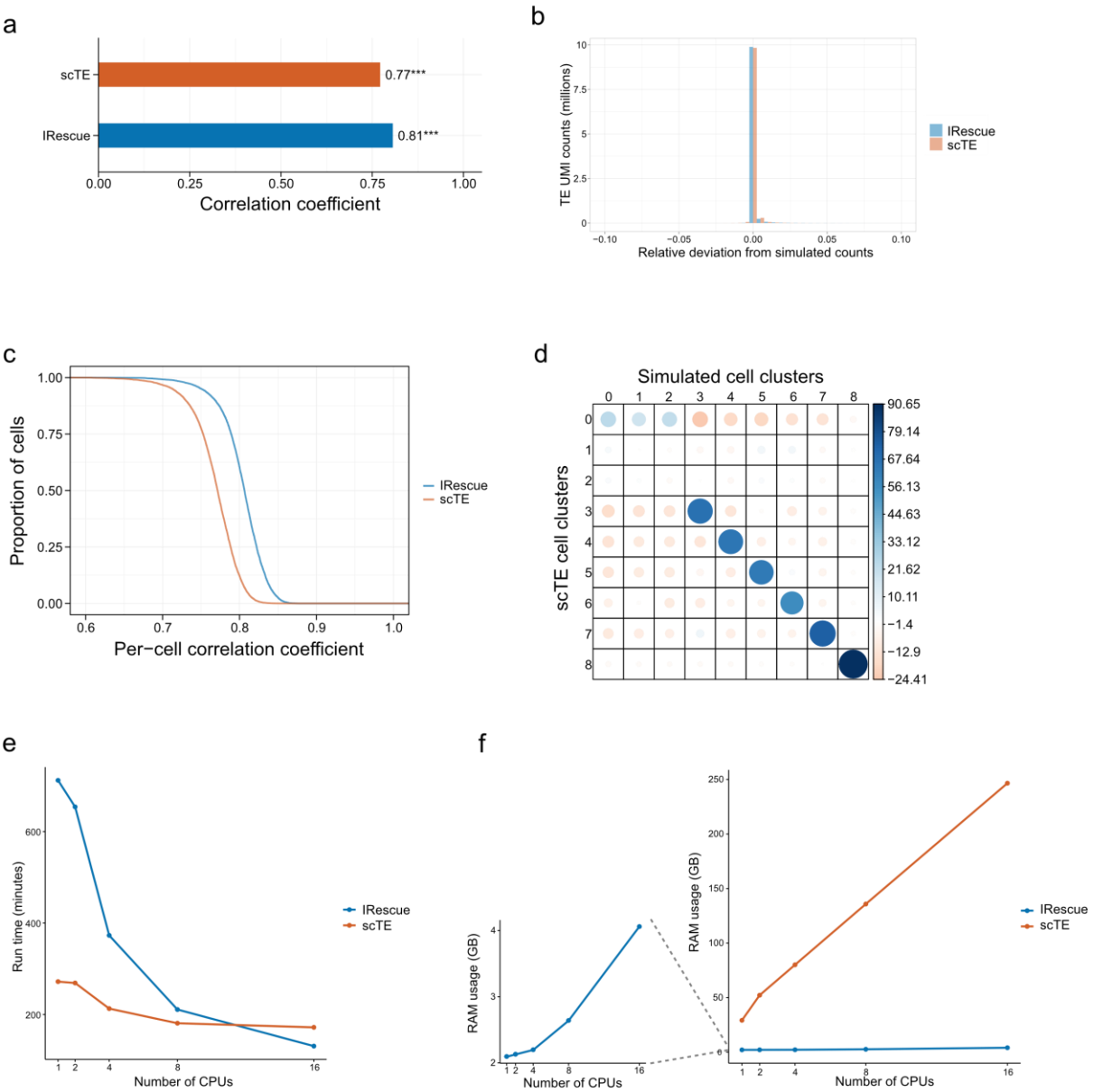
b


Inputs
Reads [fastq]
Genome [fasta]
Gene annotation [gtf]
TE annotation [BED]
scRNA-seq aligner (e.g. STARsolo)
Pre-processing
Single cell gene count matrix [MEX]
Alignments [BAM]
Filtered cell barcodes
IRescue
TE quantification
Single cell TE count matrix [MEX]
Single cell analysis toolkit (e.g. Seurat, Scanpy)
scRNA-seq analysis
Data transformation and visualization, Gene/TE expression signatures

c

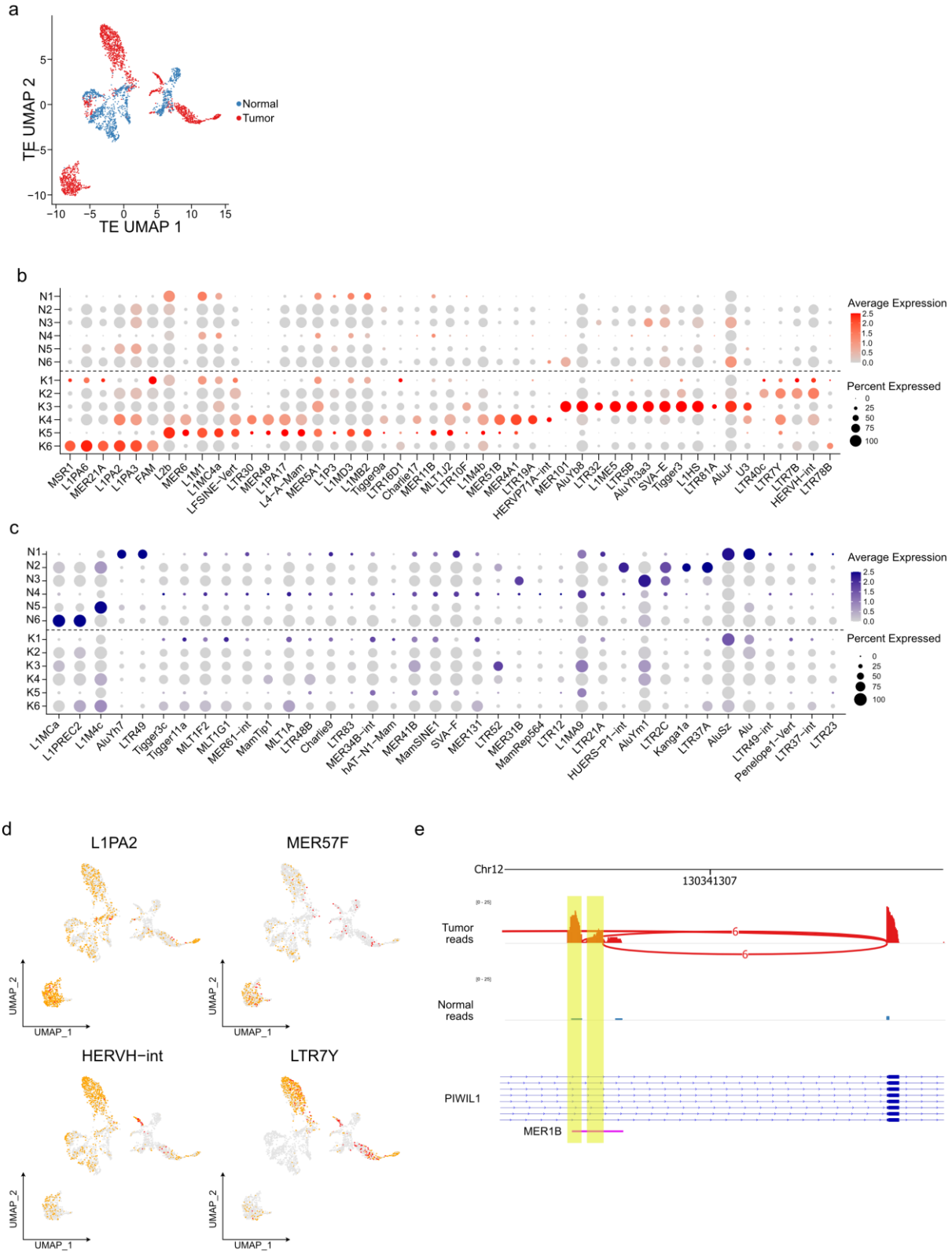

UMI sequences
1: ATGGATTGAT
1: ATGGATTGAT
2: ATGCATTGAT
3: ATGTATTGAT
1: ATGGATTGAT
2: ATGCATTGAT
3: ATGGCTTGAT
4: ATGGCTTGAG
UMI deduplication networks
|N1| = 1
N1 = N2 = N3 = {1,2,3}
|N3 ∩ N4| < |N3|
|N1 ∩ N4| < |N1|
Node (UMI)
Hub (UMI)
N = neighborhood
Contribution to TE counts
TE count +1
TE count +1
TE count +2 (number of hubs)

Supplementary figure 1. Integration of IRescue in a scRNA-seq pre-processing and analysis workflow. a. Diagram depicting a workflow for scRNA-seq data analysis of TE expression using IRescue, where boxes represent data flow (with the data format indicated in squared brackets), diamonds represent the software used and highlighted boxes represent the portion of the data flow concerning IRescue. Reads are mapped on a reference genome using a scRNA-seq splicing-aware aligner able to export the UMI and cell barcode sequences as SAM tags (e.g. STARsolo, Cell Ranger). The alignments file, the TE annotation file and, optionally, a list of filtered barcodes by the gene quantification method (e.g. the filtered barcode files exported by STARsolo or Cell Ranger) are processed with IRescue to obtain a single cell TE count matrix in Matrix Exchange format (MEX), compliant with the 10x Genomics Cell Ranger output, and therefore readable by common single cell downstream analysis toolkits such as Seurat or Scanpy. **b.** Three exemplary UMI deduplication networks, where the contribution of the count of the equivalence class (EC count) is indicated at the bottom. From left to right: a UMI with no duplicates generates a graph of one node only; three duplicate UMIs with one-mismatch distance generate a graph with three nodes connected to each other; four UMIs with up to two-mismatch distance generate a complex graph whose contribution to the TE count is solved by identifying the graph's hubs (yellow nodes). **c.** Schematic representation of possible configurations of UMIs (narrow rectangles) mapping over TEs (hot pink boxes) and gene's exons (blue boxes). IRescue retrieve all the UMIs spanning TEs (violet and green). In this example, the usage of an alternative splice donor site causes UMIs to span between an exon and an intronic TE (violet). UMIs mapping completely within TEs are ambiguously mapped on multiple TE loci (green). UMIs mapping only on exons (yellow) are not used for TE quantification.



Supplementary figure 2. Accuracy assessments and benchmarking. a. Spearman correlation coefficients of TE counts measured with IRescue or scTE against simulated counts (*** = Spearman's rank correlation test's p-value < 2.2×10−16). **b.** Distance of TE counts measured with IRescue or scTE from simulated counts, expressed as relative deviation (see methods). Negative and positive values represent under- and over-estimated counts, respectively. **c.** Reverse cumulative distribution of cells according to the Spearman correlation coefficient calculated between simulated and measured TE counts per cell. **d.** Correlation matrix plot between cell clusters inferred from scTE and simulated data, using the Louvain

algorithm implemented in Seurat v4. Pearson's correlation residuals are plotted. Each pairwise combination is shaded blue or red depending on the correlation having a positive or negative sign, respectively, and the color intensity and size of the circles are proportional to the magnitude of the correlation. p-value < 2.2×10^{-16} (chi-squared contingency table test). **e-f.** Run time of IRescue and scTE as a function of (e) number of CPUs or (f) memory usage.



Supplementary figure 3. Identification of CRC cluster-specific TE subfamilies. a. UMAP representation of CRC and normal cells according to TE expression, with colors corresponding

to cell condition. **b-c.** Dot plot representing the average expression of differentially expressed TEs in tumor (b) or normal (c) TE expression clusters (log2 fold change > 1 and adjusted p-value < 0.05, Wilcoxon rank sum test). The dot size is proportional to the percentage of expressing cells. **d.** UMAP representation of CRC and normal cells according to TE expression. Color scale represent the scaled expression of representative TE CRC markers. **e.** Sashimi plot representing the coverage of a MER1B-derived CRC-specific cryptic exon of PIWIL1 oncogene and the reads splitted between the cryptic and canonical exons.