Integrating long-read RNA sequencing improves locus-specific quantification of transposable element expression

Sojung Lee[1,2], Jayne A. Barbour[1], Yee Man Tam[1], Haocheng Yang[1], Yuanhua Huang[1,3], Jason W. H. Wong[1,2]

[1]School of Biomedical Sciences, University of Hong Kong, Hong Kong SAR, China.

[2]Centre for Oncology and Immunology, Hong Kong Science Park, Hong Kong SAR, China.

[3]Center for Translational Stem Cell Biology, Hong Kong Science and Technology Park, Hong Kong SAR, China

Abstract

Transposable elements (TEs) can influence human diseases by disrupting genome integrity, yet their quantification has been challenging due to the repetitive nature of these sequences across the genome. We developed LocusMasterTE, a method that integrates long-read with short-read RNA-seq to increase the accuracy of TE expression quantification. By incorporating fractional transcript per million (TPM) values from long-read sequencing data into an expectation-maximization (EM) algorithm, LocusMasterTE reassigns multi-mapped reads, enhancing accuracy in short-read-based TE quantification. Validated with simulated and human datasets, LocusMasterTE holds promise for revealing new insights into TE functions through precise quantification.

Keywords

Transposable Elements, Short-read RNA-seq quantification, Expectation-Maximization model

Background

Transposable elements (TEs) are repetitive genomic units accounting for half of the human genome [1]. TEs can also be grouped based on genetic composition and monophyletic origin: class > family > subfamily [1]. Consensus sequence within the subfamily accounts for the high similarity in sequence compositions of TEs [2]. In normal cells, TEs are firmly regulated through epigenetic control, including histone modifications, but in cancer cells, an extensive number of TEs are associated with epigenetic dysregulation [3]. With the accumulation of mutations through evolution [4], most TEs are now inactive at the transcriptional level in humans; however, recent studies have identified TEs with functional impact on health and disease [5]. Moreover, those remaining active TEs influence regulatory elements and can have substantial implications in tumorigenesis [6, 7].

Interest in the study of TEs has led to the development of tools for their quantification by short RNA-seq data. Some of these include Telescope [8], SalmonTE [9], SQuIRE [10], TEtranscripts [11], TEtools [12], and RepEnrich [13]. However, as TEs are highly

repetitive, especially within the same subfamily, there are many cases where one read from short-read RNA-seq is aligned to numerous TEs with similar or even identical sequence composition [10]. In particular, newly inserted TEs (i.e. young TEs) generally accumulate fewer polymorphisms with less unique sequences, leading to high multi-mapping rates [11]. Some tools avoid the multi-mapping issue by only focusing on TE quantification at the subfamily level. However, considering individual TE's are under epigenetic regulation, not all TEs in the same subfamily will simultaneously be up or down-regulated [14]. Other tools have utilized the Expectation-Maximization (EM) algorithm to estimate TE expression at specific locus [12]. Nevertheless, there are still numerous instances where the EM algorithm can only partially resolve multi-mapping reads due to insufficient information.

In recent years, long-read sequencing has emerged as a complementary approach for analysing genomes and transcriptomes. Due to its longer technical read length, long-read can increase the success rate in quantifying novel transcripts from repetitive regions and identifying full-length isoforms [15, 16]. However, its cost remains high. Hence, it is rarely applied to large datasets or at a depth comparable to short-read sequencing.

Taking advantage of the mappability of long-read and the depth of short-read RNA-seq data, we have developed LocusMasterTE, a long-read assisted short-read TE expression quantification method. LocusMasterTE integrates long-read information to aid the quantification of RNA expressed from individual copies of TEs in a locus-specific manner from short-read RNA-seq data. We demonstrate the benefits of LocusMasterTE over existing tools using simulated and human RNA-seq datasets.

Results

Long-read RNA-seq is superior to short-read RNA-seq at mapping young TEs

Long-read sequencing can sequence fragments over 10-100 kilobases, and even up to several megabases [17]. This greatly decreases the proportion of multi-mapping reads during alignment to TE in transcriptomics data (Fig. 1A). To examine the differences between long-read and short-read transcriptome sequencing for TE expression quantification, we used paired data generated from the human colon cancer cell line HCT116 to compare transcript quantification at the TE subfamily and individual level (detailed in the Methods).

As TE subfamily is an aggregation of individual copies, we observed good agreement between Oxford Nanopore Technologies (ONT) long-read and short-read (R^2=0.83, Fig. 1B); however, there was a more substantial difference at locus-specific copy level (R^2=0.43). For comparison, the correlation at coding genes is R^2=0.55 (Supplementary Fig. 1A). To account for read depth, we downsampled the short-read sample to match the number of mapped bases and conducted the same comparisons. We consistently observe more conformity at the subfamily level compared with the locus-specific TE level (Supplementary Fig. 1B).

TEs that have undergone transposition in the genome more recently (i.e. evolutionarily younger TEs) generally have fewer polymorphisms, rendering individual copies more challenging to distinguish and map uniquely. To correctly quantify young TEs using short-reads, TE quantification tools have applied expectation-maximization (EM)

algorithm to reallocate multi-mapped reads based on an estimated probability. In theory, EM algorithm should identify the most likely genomic locus for each multi-mapping read. However, we identified 6% of cases where the EM algorithm could not identify a unique genomic location due to insufficient information (Fig. 1C). The estimated age (quantified by % divergence from a consensus ancestral sequence) of these ambiguously assigned TEs is significantly younger (i.e. lower % divergence) than other quantified (14.5% versus 19.0%; p-value < 0.0001, Student's t-test, Fig. 1D), suggesting that existing EM-based methods are more prone to misquantify younger TEs.

Given that the longer read length in long-read sequencing leads to lower multi-mapping rates, we observed that the divergence of TEs quantified by long-read was lower than those captured by short-read (15.8% vs 18.7%; p-value < 0.0001, Student's t-test). The difference was increased when we compared a group of TEs exclusively quantified by long-read (median = 15.2%) and TEs only quantified by short-read (median = 19.5%; p-value < 0.0001, Student's t-test) (Fig. 1E). An example of a TE repetitive element mappable only by long-read is illustrated in Fig 1F (L1MC5a; chr7:152,051,230-152,051,364).

LocusMasterTE along-read assisted short-read TE quantification method

Given the benefits of long-read in studying TEs, we developed LocusMasterTE, which integrates long-read information to improve short-read-based TE quantification. As described above, the application of EM does not resolve all multi-mapping reads (Fig. 1C). LocusMasterTE additionally uses long-read information during the EM step, allowing improved resolution of multi-mapping (Fig. 2).

LocusMasterTE is divided into 4 stages: 1) Alignment with annotations of reads to known transcripts, including TEs, 2) Assigning initial proportions, 3) Long-read aided EM model, and 4) Final calculation of read counts. LocusMasterTE requires three inputs: a short-read sample in BAM format, a GTF file, and, most importantly, Transcript Per Million (TPM) counts from a long-read sample. As LocusMasterTE was built to quantify individual copies of TEs, a GTF containing the coordinates of individual TE copies is required. During the alignment and annotation stage, reads are divided based on the number of mapped TEs. Uniquely mapped read refers to a read fragment originating from one locus. When a read is aligned to multiple TE copies, the read is classified as a multi-mapped read (Fig. 2A). As with Telescope [8], LocusMasterTE represents the proportion of read fragments as a Bayesian mixture model and utilizes the EM algorithm to estimate parameters.

In reference to Telescope [8], LocusMasterTE uses two parameters, theta (θ) and pi (π), for the EM model, each representing the proportion of fragments originating from the transcript and the proportion of non-unique fragments to be reassigned respectively (Fig. 2B). The main innovation of LocusMasterTE lies in the integration of long-read TPM fractions (Details in Methods). For every EM iteration, long-read information is integrated into calculating the maximum *a posteriori* (MAP) for π and θ, and the updated MAPs are used for subsequent E-step (Fig. 2C). In order to demonstrate LocusMasterTE's capability to minimize cases where the EM algorithm is unable to identify a single genomic position for each read that maps to multiple locations, we conducted a similar analysis as depicted in Fig 1C. The occurrences in which the EM algorithm failed to identify a single locus were reduced to 1% with the use of

LocusMasterTE (Supplementary Fig. 2), indicating the benefit of incorporating long-read RNA-seq information.

Although data from any long-read sample can be used in LocusMasterTE, the best accuracy would be achieved using matched cell or tissue types between long-read and short-read. Nevertheless, we recognize that long-read data may not always be available for the desired cell type, thus, LocusMasterTE introduces a weight parameter (--*long_read_weight*), allowing users to manage the contribution of long-read information. Lower weights are recommended with distant tissue types, but optimal weights vary between samples. We conducted tests with various weight (--*long_read_weight*) values from 0.01 to 1.5 on our in-house HCT116 short-read cell line sample along with a K562 long-read sample from the SG-NEx study [18]. Our findings revealed that values below 0.5 and above 1.0 (--*long_read_weight*) resulted in lower accuracy (Supplementary Fig. 3). When an unmatched long-read sample is used, it is advisable to use a weight (--long_read_weight) smaller than 1.0. Otherwise, in most cases, a weight (--*long_read_weight*) of 1.0 is recommended (default). At the conclusion of the EM optimization, LocusMasterTE provides an updated alignment with finalized counts (Fig. 2D).

Benchmarking LocusMasterTE on simulated data

To demonstrate the accuracy of LocusMasterTE, we simulated short-read RNA-seq datasets based on a long-read RNA-seq sample (Details in Methods), enabling the use of the long-read data as the ground truth. Using the simulated short-reads, we compared the accuracy of LocusMasterTE for TE expression quantification with seven approaches: 1) Unique counts (Fig. 3A), 2) featureCounts [18] (Fig. 3B), 3) RSEM [19] (Fig. 3C), 4)

Telescope [8] (Fig. 3D), 5) SQuIRE [10] (Fig. 3E), 6) SalmonTE [9] (Fig. 3F), 7) LocusMasterTE (Fig. 3G). "Unique counts" was directly provided by Telescope [8]. For fair comparisons, salmonTE [9] was modified to output TE quantifies at individual loci, as it was originally designed to calculate TE at the subfamily level. Moreover, we allowed mult-mapping read (-*M*) and overlapping features (-*O*) to be counted in featureCounts [18].

Prior to comparing TE counts, we tested the quality of simulated short-read using coding genes (Supplementary Fig. 4). Out of 19120, 18465 coding genes were consistently captured between long-read and simulated short-read (TP = 13229, TN = 5236, FP = 120, FN = 535). LocusMasterTE had no falsely detected TEs while having the highest number of correctly detected TEs (TP = 45686). It had the highest precision among the seven approaches (Precision: 1.0, Recall: 0.86, F1 score: 0.92; Fig. 3H). FeatureCounts [18] showed relatively low precision (Precision: 0.46, Fig. 3H). As featureCounts [18] reports all multi-mapping reads without the EM algorithm, low precision in featureCounts demonstrates the importance of the EM step when assigning multi-mapping reads. SQuIRE [10] captures a similar amount of TEs as LocusMasterTE, but has a higher false positive rate, leading to low precision (Precision: 0.49, Fig. 3H). As LocusMasterTE was built upon Telescope [8], we carefully compared the performance of the two tools. LocusMasterTE achieved much higher TP and TN with lower FP and FN values, outperforming Telescope [8] (Fig. 3D, 3G). Based on this simulated dataset, LocusMasterTE is demonstrated to be the most accurate among the seven approaches.

To ensure compatibility with both ONT and PacBio Iso-Seq long-read RNA sequencing inputs, we conducted a performance comparison of LocusMasterTE using our simulated

short-read RNA-seq from these two platforms. Since PacBio Iso-Seq primarily aims to identify novel transcripts, we utilized Talon [22] for quantification. While there is a slight variation in quantified TE expression between ONT and PacBio, we found that 94.7% (N=85467 of 97746) of individual TEs quantified are consistently captured by both techniques (Supplementary Fig. 5A). Additionally, we observed a strong positive correlation between the two long-read techniques (spearman correlation estimates (ρ) = 0.9, Supplementary Fig. 5B), indicating that both long-read approaches can serve as suitable inputs for LocusMasterTE.

TEs uniquely quantified by LocusMasterTE are evolutionarily younger

To demonstrate the advantage of using LocusMasterTE to study TE expression, we used cell line samples with matched long and short-read RNA sequencing data as input for LocusMasterTE. This included one in-house dataset from HCT116 and five human cancer cell lines (A549, HCT116, HepG2, K562, and MCF-7) from the SG-Nex [20].

We examined TEs that showed expression changes between LocusMasterTE and Telescope [8] and found that these were generally evolutionarily younger TEs compared with those that were unaffected by the choice of method. This suggests that LocusMasterTE is indeed correcting the expression of younger TEs that are generally difficult to quantify (Fig. 4A). TE counts generated by Telescope [8] and LocusMasterTE were then compared with long-read results to confirm the successful integration of long-read in cell line samples. We observed an increased correlation in TE expression with long-read results using LocusMasterTE (R^2 = 0.39) compared with Telescope (R^2 = 0.35) using the inhouse HCT116 dataset [8] (Fig. 4B). Due to the difference in average sequencing depth between in-house and SG-NEx samples (Supplementary Fig. 6), there

were slight variations in correlation values with long-read. Nevertheless, the same trend was consistently observable in SG-NEx across all cell types. (Fig. 4C ; p-value=0.0062, Paired t-test, Supplementary Fig. 7).

TEs solely quantified by LocusMasterTE are marked by more active and less repressive chromatin

Histone modifications regulate gene expression, including those of TEs. We selected TEs that were differentially quantified by LocusMasterTE and Telescope [8], grouping them based on whether a TE had a higher expression value as quantified by the respective method (detailed in Methods). We then examined their associations with five histone marks: three active (H3K4me3, H3K9ac, H3K27ac) and two repressive (H3K27me3, H3K9me3). To avoid bias in the histone ChIP-seq data, we selected TE regions with a mappability score greater than 0.25, finding no significant difference between the two groups (median TEs with higher expression by Telescope) = 0.5006275 vs median (TEs with higher expression by LocusMasterTE) = 0.505051, p = 0.1661, Wilcoxon signed-rank test, Supplementary Fig. 8A). Between two groups of TEs, TEs highly captured by LocusMasterTE had less enrichment for the two repressive marks (Fig. 4D), and more enrichment for all three active marks (Fig. 4E) The same trends were observed across all cell lines (Supplementary Fig. 8B). This suggests that LocusMasterTE improves TE qualification as it is more consistent with independent epigenetic data.

Presence of RNA editing validates novel transcripts identified by LocusMasterTE

To validate that reads mapping to TEs that LocusMasterTE reassigned are genuinely TE-derived, we utilized samples from the TCGA-COAD cohort to study RNA editing events

at TEs. RNA editing refers to the post-transcriptional modification of RNA sequences by specific enzymes. Recent studies reveal that Alu elements account for more than 99% of human RNA editing events by forming double-stranded RNA that undergoes A-to-I editing [22, 23]. The TCGA-COAD cohort (N = 287) was used to ensure sufficient sequence depth for RNA editing detection.

We derived the reference RNA editing trinucleotide signature from MDA5-protected Alu elements from 5-azacytidine-treated ADAR wild-type samples [28] and compared them with three groups from our analysis: (1) coding genes, (2) TEs detected by both Telescope [8] and LocusMasterTE, labeled as "common", and (3) TEs exclusively captured by LocusMasterTE, labeled as "only in LocusMasterTE". The RNA editing profile of TE classified as “common” and “only in LocusMasterTE” showed a similar pattern with the reference sites, with peaks at CTA > CCA, CTG > CCG, and GTG > GCG (Fig. 5A). This is supported by a high cosine similarity (Fig. 5B), but with a clear difference from coding genes that generally show much lower editing levels. This suggests that reads reassigned by LocusMasterTE are likely to be from TEs-derived RNA which are subjected to RNA editing.

Discussion

Transposable elements play essential roles in cancer and complex biological diseases in humans [1]. However, due to the repetitive nature of TEs sequences, it has been challenging to align and quantify TEs accurately. Since long-read has an advantage over short-read in mappability due to read length, we developed LocusMasterTE to integrate the long-read results using the EM algorithm. Including long-read results enables the EM

algorithm to assign reads more correctly than conventional short-read TE quantification tools. In particular, LocusMasterTE resolves almost all cases where read assignment remains ambiguous despite the use of the EM algorithm. Also, LocusMasterTE allows users to use any long-read regardless of cell or tissue type.

The accuracy of LocusMasterTE was first illustrated using a simulated short-read dataset. Among seven TE quantification approaches, LocusMasterTE successfully achieved the highest recall and precision scores. Additionally, our study reveals the advantage of using LocusMasterTE in studying evolutionarily young TEs. As newly inserted TEs achieve fewer polymorphisms, many tools struggle to capture young elements correctly. Based on a comparison with Telescope [8], TEs uniquely quantified by LocusMasterTE were generally younger. TEs quantified by LocusMasterTE were also more consistent with active and repressive histone marks quantified independently by ChIP-seq [21]. Finally, we leverage the presence of RNA editing to validate that reassigned reads by LocusMaster TE are indeed TE-derived.

While we believe LocusMasterTE is robust and accurate for TE quantification, it does have a few limitations. Since LocusMasterTE relies on long-read data, having long-read sequencing information is essential. However, considering the higher cost of long-read sequencing compared to other techniques, not all samples have both short-read and long-read data simultaneously, or at least not with sufficient long-read sequencing depth. Although LocusMasterTE allows users to input long-read data from any type of sample, the best performance is achieved when using long-reads from a matched sample or cell type. When employing the same reference long-read sample and running LocusMasterTE

on different short-read samples, the TE counts generated by LocusMasterTE may include some estimated counts. Secondly, the performance of LocusMasterTE depends on the coverage of long-read data. With deeper sequencing depth, LocusMasterTE quantifies TEs more accurately. Despite these limitations, LocusMasterTE achieves the most precise TE quantification among available tools and effectively identifies previously undetected active TEs.

Conclusion

Leveraging the advantages of long-read sequencing, we developed LocusMasterTE, which incorporates long-read RNA-seq data as a prior term to resolve multi-mapped reads in short-read RNA-seq datasets. This approach is adaptable to both PacBio and ONT long-read inputs, allowing users to adjust the weighting of long-read information. Our analysis of cell lines demonstrated that LocusMasterTE effectively captures young TEs, with their activity supported by histone marks. Moreover, in the TCGA-COAD cohort, TEs identified by LocusMasterTE exhibited clear RNA editing signatures, highlighting its potential for uncovering novel insights into TE activities in patient samples.

Methods

Datasets Used in Study

We used three main datasets for our study: 1) Simulated short-read data from an in-house HCT116. 2) Paired short-read and long-read (ONT) RNA sequencing data from in-house HCT116 and five different cell line samples from the SG-Nex study [18]. 3) TCGA-

COAD tumor samples: we utilized a total of 287 TCGA-COAD tumor samples. In-house HCT116 long-read was used as input for LocusMasterTE.

Generation of HCT116 RNA sequencing data

The cells with the plate number HCT116 were cultured in McCoy's 5A media supplemented with 10% FBS and then seeded into 35 cm dishes. When they reached 80% confluence, RNA was extracted using the Qiagen RNeasy Mini kit (74104), following the manufacturer's directions.

For short-read RNA sequencing, the samples were sent to Novogene Co., Ltd (Beijing) for library preparation and Illumina sequencing. The process involved enriching the mRNA with oligo (dT) beads, fragmenting it randomly, synthesizing cDNA from the fragmented mRNA using random hexamers as primers, and then subjecting it to second-strand synthesis, end-pair, dA-tailing, and ligation of sequencing adapters. The sequencing was performed on an Illumina Novaseq 6000 instrument.

For long-read RNA sequencing, 50 ng of the RNA was used for library preparation using the Oxford Nanopore cDNA-PCR barcoding kit (Oxford Nanopore, SQK-PCB109), following the manufacturer's protocol assuming an average transcript length of 1.5 Kb. Then, 100 fmol of the library was loaded on one flow cell (Oxford Nanopore, FLO-MINSP6), which was sequenced on the MinION Sequencing Device (Oxford Nanopore, MIN-101B).

Generation of Simulated short-read RNA sequencing

To benchmark LocusMasterTE, we generated a simulated short-read from long-read RNA sequencing using the read simulator Spanki [42]. We used in-house HCT116 ONT

as input for the simulation. The simulation was only performed on chromosomes 1-22, X, and Y. Spanki allows users to input a transcript coverage file, and we inputted HCT116 ONT long-read TPM values for each TE as coverage [42]. To match the number of bases in the generated FASTQ [41] file for simulated short-read and long-read, we first obtained the simulated FASTQ file without mapping the number of bases. Let the number of bases mapped in long-read FASTQ as BaseMapped_{long} and the number of bases mapped in simulated paired-end short-read FASTQ without mapping coverages as $\text{BaseMapped}_{\text{sim.short}}$. Then we calculated the ratio between the two as follows:

$$\frac{\text{BaseMapped}_{\text{long}}}{\text{BaseMapped}_{\text{sim.short}}} = \frac{2277300856}{95915952} \approx 23$$

Based on the calculated values, we multiplied 23 in long-read TPM values, and updated TPM values were used as coverage input for simulation. Default values 76 and 200 were used for read length and fragment size, respectively.

Adding Long-read information into the prior distribution of EM algorithm

In this study, LocusMasterTE was developed based on the fragment reassignment mixture model, Telescope [8], to enable locus-specific quantification of transposable elements. Specifically, LocusMasterTE utilizes an expectation maximization (EM) algorithm [22] to calculate an optimal probability for each read assignment. The essential innovation of LocusMasterTE lies in the EM model, where long-read information is used as prior term. Long-read sequencing data, ideally from a related cell type, is first aligned to a reference genome such as hg38 using Minimap2 [23]. Following alignment, TE quantification of long-read is performed using featureCounts [18] with a GTF file, such

as one generated by RepeatMasker [24], containing individual TE transcripts $T = \{t_1, t_2, \dots, t_K\}$, where K is the number of transcripts. Raw counts for each TE generated by featureCounts [18] is converted to Transcripts Per Million (TPM), such that each transcript has a TPM value, t_i^L, (i.e. $T^L = \{t_1^L, t_2^L, \dots, t_K^L\}$). To ensure compatibility during model integration, TPM counts are converted to fractions($T^{Lf} = \{t_1^{Lf}, t_2^{Lf}, \dots, t_K^{Lf}\}$). As sequence composition is very similar within subfamily, TEs are normalized within in subfamily.

$$t_i^{Lf} = \frac{t_i^L}{\sum_{i=1}^{K} t_i^L \in subT}$$

where *subT* is the set of TEs from the subfamily of t_i.

Subsequently, LocusMasterTE follows the conventional EM model with expectation (E-step) and maximization (M-step), where an integration of t_i^{Lf} occurs. Finally, let a set of indicators $Y = \{y_1, y_2, \dots y_N\}$.

$$y_i = 1 \text{ in multi-mapped cases}$$
$$y_i = 0 \text{ otherwise}$$

As introduced by Telescope [8], $\pi = \{\pi_1, \pi_2, \dots, \pi_K\}$ and $\theta = \{\theta_1, \theta_2, \dots, \theta_K\}$ are two sets of parameters: representing the proportion of observed fragments originating from each transcript, and similar reasignment parameters for multi-mapped reads, respectively [14]. In the initialization stage, LocusMasterTE assigns equal weights to all reads. In E-step, LocusMasterTE calculates the posterior probability of partial assignment weights for all fragments, denoted as $E\left[t_i^{S_a}\right]$ for the posterior probability of assigning read S_a to transcript i. Next, in M-step, based on the estimated probability from the E-step,

maximum *a posteriori* (MAP) is estimated for model parameters. Here, LocusMasterTE includes the T^{Lf} term as the prior distribution. LocusMasterTE has an useful option, (*--long_read_weight)*, allowing users to freely adjust the weights of the long-read during integration. Let w be the value for (*--long_read_weight*) option.

$$\hat{\pi}_i = \frac{\sum_{a=1}^{N} E\left[t_i^{sa}\right] + (w \cdot \sum_{i=1}^{K} \sum_{a=1}^{N} E\left[t_i^{sa}\right]) \cdot t_i^{Lf}}{\sum_{i=1}^{K} \sum_{a=1}^{N} E\left[t_i^{sa}\right] + \sum_{i=1}^{K} (w \cdot \sum_{i=1}^{K} \sum_{a=1}^{N} E\left[t_i^{sa}\right]) \cdot t_i^{Lf}}$$

$$\hat{\theta}_i = \frac{\sum_{a=1}^{N} E\left[t_i^{sa}\right] y_a + (w \cdot \sum_{a=1}^{N} y_a) \cdot t_i^{Lf}}{\sum_{a=1}^{N} y_a + \sum_{a=1}^{N} (w \cdot \sum_{a=1}^{N} y_a) \cdot t_i^{Lf}}$$

From the mixture model provided by Telescope [8], introducing a new parameter T^{Lf} during the EM algorithm, changing the parameters π and θ, which consequently affects overall mixture weights in the model [8]. Using the updated MAPs of two parameters, LocusMasterTE runs E-steps and M-steps iteratively until the changes in two parameter estimates are less than the specific level, which is usually set to $\epsilon < 0.001$ [25]. All other formulas are in accordance with the Telescope [8].

Other essential features in LocusMasterTE

1. Due to differences in sequencing techniques, some TEs can only be detected by either short-read sequencing or long-read sequencing. However, because LocusMasterTE uses long-read TPM counts directly and updates parameters through multiplication, if the quantified TPM counts from long-read sequencing are 0, then regardless of how many TEs are captured by short-read sequencing, the final counts will also become 0. To address this issue and rescue TEs that are only quantified by short-read sequencing, we have introduced the (*--rescue_short*) argument. This argument replaces 0 with a small constant in the long-read TPM

count. The default value is set to 0, but users have the flexibility to input any small constant (i.e. $1e^{-50}$) based on their preferences.

2. We have added an extra option named "long_read" in the (--reassign_mode) argument, originally introduced by Telescope [8]. Telescope [8]."long_read" option re-distributes counts based on the weighted mean. Assume after running EM, TE_A, TE_B, and TE_C are having same values $TE_A^c = TE_B^c = TE_C^c$ for read R_1. Then, the final count for TE_A represented TE_A^{fc}, is formulated below.

$$TE_A^{fc} = (TE_A^c + TE_B^c + TE_C^c) \cdot \frac{t_A^L}{t_A^L + t_B^L + t_C^L}$$

Gene and TE quantification

We obtained the coding gene Gencode annotation of the hg38 version 42 Basic GTF file [36] from the UCSC table browser using MANE annotated gene filters [37]. The GTF file for TE quantification was created using RepeatMasker [21] for Genome Reference Consortium Human Build 38 (GRCh38) [38]. We used Minimap2 [32] with GRCh38 fasta to align ONT. For Pacbio Isoseq, we used the isoseq3 tool developed by PacificBiosciences. Specifically, *CCS* was used to compute circular consensus sequences, and *lima* and *isoseq refine* were applied to select and process *CCS*. Minimap2 [32] was used for alignment, and Talon [37] was used for quantification. For short-read alignment, we used STAR [39] as the aligner. TE quantification was performed using LocusMasterTE for all analyses, except for long-read and short-read comparisons in Figure 1. We used the parameters below for LocusMasterTE.

- (*--prior_change*): both, meaning changing both π and θ
- (*--long_read_weight*): 1
- (*--rescue_short*): $1e^{-50}$
- (*--reassign_mode*): long_read

In Figure 1, featureCounts was used to assign genomic features to reads for both long-read and short-read [33]. Additionally, since LocusMasterTE originated from Telescope, several comparisons were made between Telescope [8] and LocusMasterTE. For each TE, two values were generated: counts from Telescope [8] and counts from LocusMasterTE, $TE^{c.tel}$ and $TE^{c.loc}$ respectively. Throughout the study, Transcripts Per Kilobase Million (TPM) was used instead of raw counts, with the sum of coding genes used for the TPM library size.

Comparative analysis on proportion of multi-mapping reads

First, for short-read RNA-seq, the STAR aligner was used to align the reads to the entire genome [39]. After that, bedtools was used to filter out regions containing basic genes [35]. Then, the remaining reads were remapped using BWA [40], and regions that overlapped with TEs were selected. For long-read RNA-seq, minimap2 was initially used for the alignment [32]. After that, unmapped reads were remapped using BWA, and reads corresponding to TEs were extracted. The number of multi-mapping reads and uniquely mapped reads was calculated based on the "FLAG" tag in the SAM file [41].

Comparison with conventional tools

To compare the accuracy of LocusMasterTE, we quantified TEs in simulated short-read RNA sequencing using Telescope [8], RSEM[19], SQuIRE [10], FeatureCounts [18], and SalmonTE [9]. To generate an index for SalmonTE, we used a fasta file consisting of TE individual sequences, which was generated by a locus-specific TE expression study [14]. We used STAR for LocusMasterTE and other tools like Telescope [8], FeatureCounts [18],and SQuIRE [10]. Bowtie2 [43] was used as an aligner for RSEM [19]. As the simulated short-read was generated from a long-read RNA sequencing sample, we considered quantified TE counts from the long-read as the ground truth. We only considered TEs captured by at least one approach, removing unsuccessfully simulated TEs. We calculated accuracy and recall values compared with the ground truth based on the confusion matrix generated for each tool.

Age analysis in cell-line samples

We acquired % divergence (% divergence from ancestral sequences) of TE individuals from the RepeatMasker table at the UCSC database.

TEs were divided into two groups. TEs having the same counts between Telescope [8] and LocusMasterTE were considered as “originally correctly measured TEs”. For TEs quantified differently between Telescope [8] and LocusMasterTE were defined as “corrected by LocusMasterTE”, because those TEs were either over-estimated or under-estimated in Telescope [8].

- $TE^{\text{c.loc}} = TE^{\text{c.tel}}$: TEs quantified same in both tools
- $TE^{\text{c.loc}} \neq TE^{\text{c.tel:}}$: TEs quantified differently

Student T-test was conducted between two groups.

Histone marker analysis in cell-line

Five histone modifications representing active (H3K4me3, H3K27ac, and H3K9ac) and inactive (H3K27me3 and H3K9me1) chromatin were used for histone analysis. K50 umap bigwig [44] was used to extract mappability over TE regions. Considering the presence of mappability bias between two groups, TE regions with mappability higher than 0.25 were selected. Fold change over control BigWig files from ChIPseq data of the five histone marks for each cell line were downloaded from ENCODE project [31]. TEs were re-divided into two groups. BigWigAverageOverBed was used to calculate average fold change over control signal in two groups of TEs.

- $TE^{\text{c.loc}} > TE^{\text{c.tel}}$: TEs having higher expression in LocusMasterTE
- $TE^{\text{c.loc}} < TE^{\text{c.tel}}$: TEs having higher expression in Telescope [8]

For better comparison, TEs with non-trivial differences between Telescope [8] and LocusMasterTE $\left|TE^{\text{c.loc}} - TE^{\text{c.tel}}\right| > 1$ were selected. Then, we used computeMatrix from deepTools [26] to calculate scores.

RNA Editing events in TCGA-COAD

TEs were first grouped into two as follows:

1. $TE^{\text{c.loc}} \neq 0 \wedge TE^{\text{c.tel}} = 0$: TEs exclusively measured by LocusMasterTE
2. $TE^{\text{c.loc}} \neq 0 \wedge TE^{\text{c.tel}} \neq 0$: TEs both measured by two tools

A total of 35000 TEs were randomly selected for the two groups. In addition, 35000 coding genes were chosen for comparison. The ADAR-WT-MDA5-5aza sample from an

epigenetic therapy study in inverted SINEs and ADAR1 [27] was used as the "reference" sample to extract expected editing signatures. The same GRCh38 fasta and genomic location files mentioned above [38] were used, and bam-readcount was employed to gather information at specific nucleotide positions [50]. Germline SNPs overlapping with the gnomAD database [51] were filtered out using bcftools [52]. Alu regions were used for all groups except for the coding genes group. The R package MutationalPatterns was used to illustrate mutational profiles for all groups in trinucleotide contexts [53]. Additionally, we conducted a series of cosine similarity tests using the *lsa* R package [54] to compare all groups with the TCGA-COAD samples [55].

Declarations

Availability of data and code

Raw RNA sequencing datasets for the TCGA-COAD study can be downloaded from the TCGA data portal [https://portal.gdc.cancer.gov/]. Our study used processed data for RNA sequencing from UCSC Xena [http://xena.ucsc.edu/]. Histone datasets and HCT116 pacbio isoseq data are retrieved from ENCODE [https://www.encodeproject.org/]. Short-read and ONT RNA sequencing datasets for SG-NEx are available through their GitHub [https://github.com/GoekeLab/sg-nex-data]. Inhouse HCT116 long-read and short-read are deposited at Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo)] under GSE225377 and GSE225380 respectively. We analyzed the ADAR-WT-MDA5-5aza sample from the accession code GSE145639.

LocusMasterTE is an open-source tool. Code for LocusMasterTE is available at the GitHub repository: https://github.com/jasonwong-lab/LocusMasterTE. Scripts used to make main figures, and supplementary information files are also available at the GitHub repository: https://github.com/jasonwong-lab/LocusMasterTE_paper

Competing Interests

All authors have no competing interests to declare.

Acknowledgements

Funding: This work was supported by funds from Centre for Oncology and Immunology under the Health@InnoHK Initiative funded by the Innovation and Technology Commission, the Government of Hong Kong SAR, China (J.W.H.W).

References

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS, et al: **Ten things you should know about transposable elements.** *Genome Biol* 2018, **19:**199.
2. Jurka J, Smith T: **A fundamental division in the Alu family of repeated sequences.** *Proc Natl Acad Sci U S A* 1988, **85:**4775-4778.
3. Feinberg AP, Vogelstein B: **Hypomethylation distinguishes genes of some human cancers from their normal counterparts.** *Nature* 1983, **301:**89-92.
4. Piegu B, Bire S, Arensburger P, Bigot Y: **A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity.** *Mol Phylogenet Evol* 2015, **86:**90-109.

5. Bhat A, Ghatage T, Bhan S, Lahane GP, Dhar A, Kumar R, Pandita RK, Bhat KM, Ramos KS, Pandita TK: **Role of Transposable Elements in Genome Stability: Implications for Health and Disease.** *Int J Mol Sci* 2022, **23**.

6. Vidal E, Sayols S, Moran S, Guillaumet-Adkins A, Schroeder MP, Royo R, Orozco M, Gut M, Gut I, Lopez-Bigas N, et al: **A DNA methylation map of human cancer at single base-pair resolution.** *Oncogene* 2017, **36:**5648-5657.

7. Chuong EB, Elde NC, Feschotte C: **Regulatory activities of transposable elements: from conflicts to benefits.** *Nat Rev Genet* 2017, **18:**71-86.

8. Bendall ML, de Mulder M, Iniguez LP, Lecanda-Sanchez A, Perez-Losada M, Ostrowski MA, Jones RB, Mulder LCF, Reyes-Teran G, Crandall KA, et al: **Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression.** *PLoS Comput Biol* 2019, **15:**e1006453.

9. Jeong HH, Yalamanchili HK, Guo C, Shulman JM, Liu Z: **An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data.** *Pac Symp Biocomput* 2018, **23:**168-179.

10. Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH: **SQuIRE reveals locus-specific regulation of interspersed repeat expression.** *Nucleic Acids Res* 2019, **47:**e27.

11. Jin Y, Tam OH, Paniagua E, Hammell M: **TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets.** *Bioinformatics* 2015, **31:**3593-3599.

12. Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C: **TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes.** *Nucleic Acids Res* 2017, **45:**e17.

13. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N: **Transcriptional landscape of repetitive elements in normal and cancer human cells.** *BMC Genomics* 2014, **15:**583.

14. Schwarz R, Koch P, Wilbrandt J, Hoffmann S: **Locus-specific expression analysis of transposable elements.** *Brief Bioinform* 2022, **23**.

15. Debladis E, Llauro C, Carpentier MC, Mirouze M, Panaud O: **Detection of active transposable elements in Arabidopsis thaliana using Oxford Nanopore Sequencing technology.** *BMC Genomics* 2017, **18:**537.

16. Shahid S, Slotkin RK: **The current revolution in transposable element biology enabled by long reads.** *Curr Opin Plant Biol* 2020, **54:**49-56.

17. Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P: **Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes.** *Trends Plant Sci* 2019, **24:**700-724.

18. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.

19. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.

20. Chen Y, Davidson NM, Wan YK, Patel H, Yao F, Low HM, Hendra C, Watten L, Sim A, Sawyer C, et al: **A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines.** *bioRxiv* 2021**:**2021.2004.2021.440736.

21. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al: **Expanded encyclopaedias of DNA elements in the human and mouse genomes.** *Nature* 2020, **583:**699-710.

22. Levitan E, Herman GT: **A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography.** *IEEE Trans Med Imaging* 1987, **6:**185-192.

23. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.

24. Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2004, **Chapter 4:**Unit 4 10.

25. Wu CFJ: **On the Convergence Properties of the EM Algorithm.** *The Annals of Statistics* 1983, **11:**95-103, 109.

26. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T: **deepTools: a flexible platform for exploring deep-sequencing data.** *Nucleic Acids Res* 2014, **42:**W187-191.

Figures

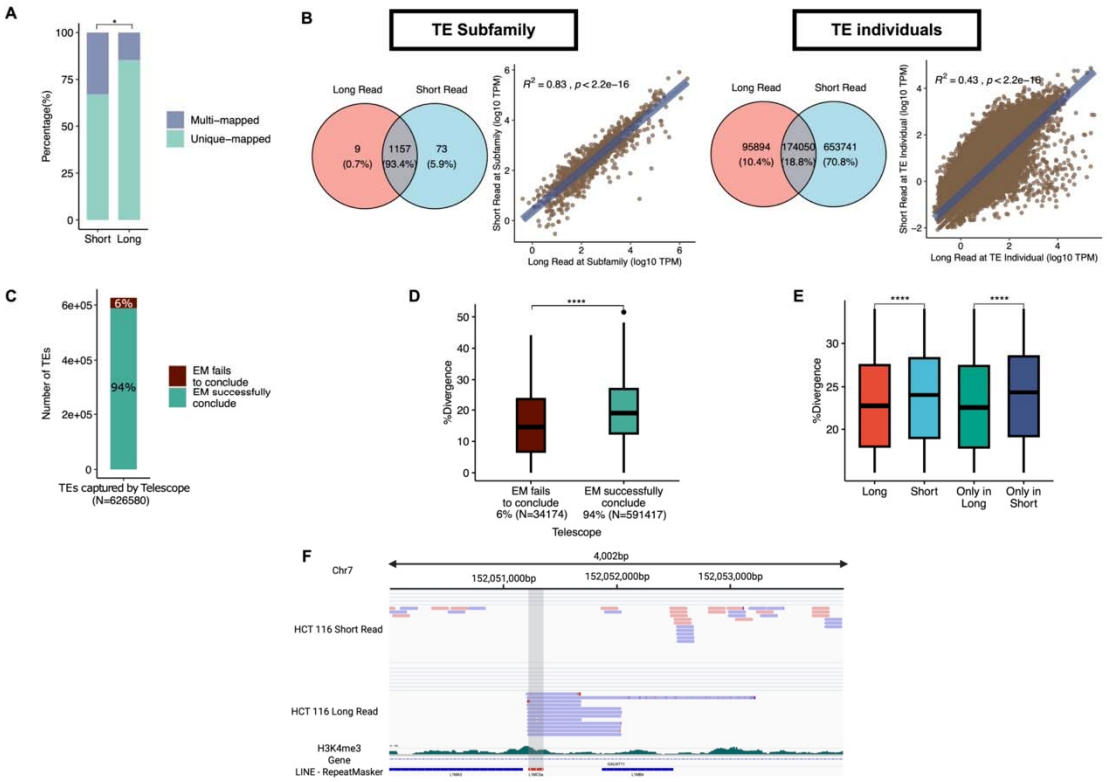

Figure 1: Comparison between long-read and short-read sequencing techniques
(A) Percentages for uniquely mapped reads and multi-mapped reads for short-read and long-read measured by BWA. (B) Left: Venn diagrams for long-read and short-read captured elements in two domains: TE subfamily and TE individuals. Right: Correlation plots between log10 TPM count for long-read and short-read in two same domains. (C) Overview of cases where EM takes part in short-read TE quantification by Telescope [8]. The total number of TEs corresponds to TEs initially aligned by Telescope [8]. Grey: TEs that were initially aligned but concluded to be not expressed after running the EM model. Green: TEs that were determined to have values as EM coverages. Red: TEs that the EM model ends up with more than two best hits, meaning it fails to conclude the genomic origin of multi-mapped reads. (D) %divergence (% divergence from the consensus sequence) in EM has more than two best hits (Brown). EM successfully concludes the best hit (Green). A lower %divergence represents younger TE. (E) Box plot showing the difference between captured TEs based on sequencing types ($p < 0.0001$). Red: TEs captured in long-read. Light blue: TEs quantified by short-read. Green: TEs are exclusively captured in long-read (Only in long). Dark blue: TEs are solely quantified by short-read (Only in Short). (E) IGV demonstrates a specific example of an advantage in

long-read for TE study. The grey area represents reads for L1MC5a; chr7:152,051,230-152,051,364.

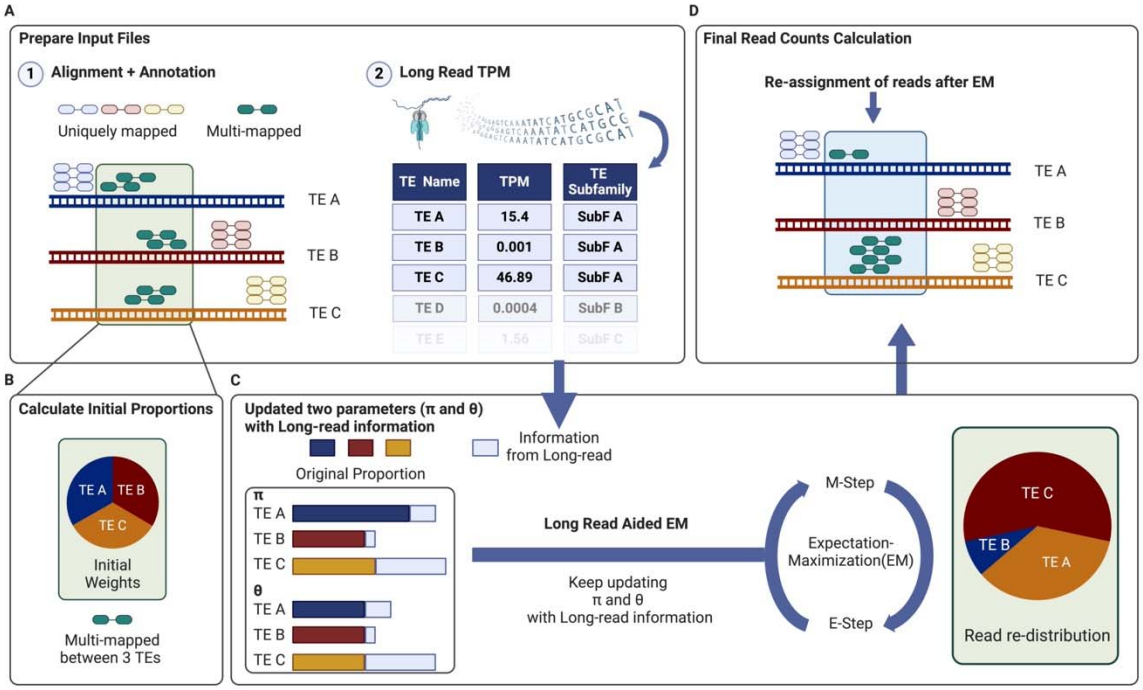

Figure 2: Schematic overview of LocusMasterTE

An example illustration for the LocusMasterTE pipeline is presented. (A) The short-read RNA sequencing data is aligned using the provided annotation file, and reads are classified into uniquely mapped and multi-mapped reads based on the alignment score. Uniquely mapped reads and multi-mapped reads are then aligned to TE A, TE B, and TE C. Long-read RNA sequencing, preferably ONT, is aligned using minimap2 [23], and TEs are quantified using featureCounts [18] with the same annotation file provided for short-read RNA sequencing. TPM counts for long-reads are calculated, and the belonging subfamily is provided for each individual TE. (B) Initially, equal weights are assigned to all three TEs - TE A, TE B, and TE C (TE A: 0.33, TE B: 0.33, TE C: 0.33). (C) Two critical parameters, pi (π) and theta (θ), represent the proportion of fragments originating from the transcript and the proportion of non-unique fragments to be reassigned, respectively. These parameters are updated with TPM fractional values from long-read RNA sequencing. During the E-step, MLE is calculated for all TEs. In the M-step, MAP estimates are updated. (D) Finally, after the EM algorithm coverages, best-aligned read counts are determined, leading to the final assignment of TE counts.

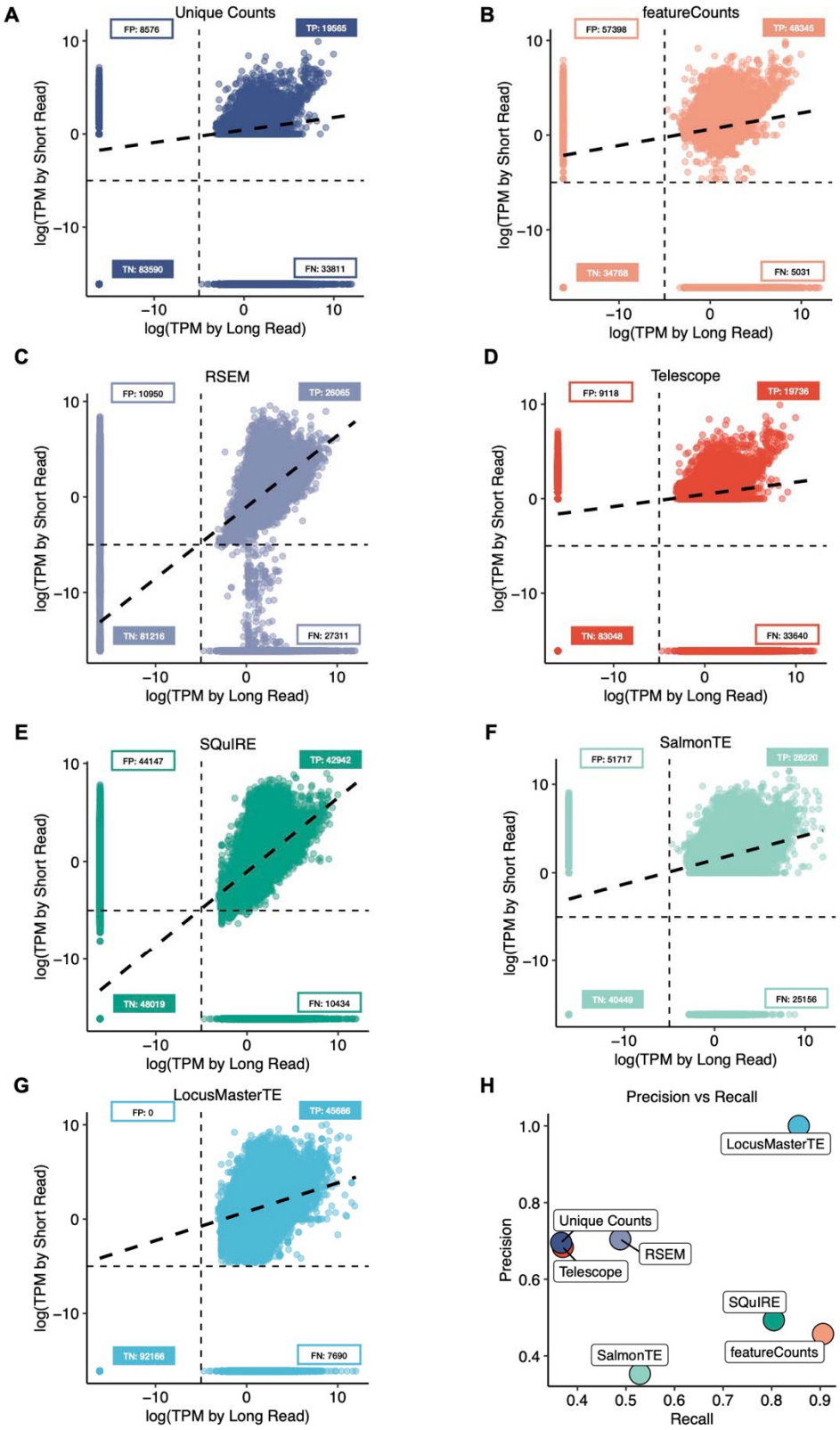

Figure 3: Assessing the accuracy of LocusMasterTE in simulated short-read

7 distinctive approaches to quantify TEs in a simulated short-read. Tested approaches are (A) unique counts provided by Telescope [8], (B) featureCounts [18], (C) RSEM [19], (D) Telescope [8], (E) SQuIRE [10], (F) salmonTE [9], and (G) LocusMasterTE. As a simulated short-read is generated from a long-read, all comparisons are conducted with a long-read TE quantification result. All quantified TE counts are converted to log10 TPM counts. For all approaches (A-G), true positive (TP), true negative (TN), false positive (FP), and false negative (FN) numbers are included. (H) Precision and recall values for seven approaches.

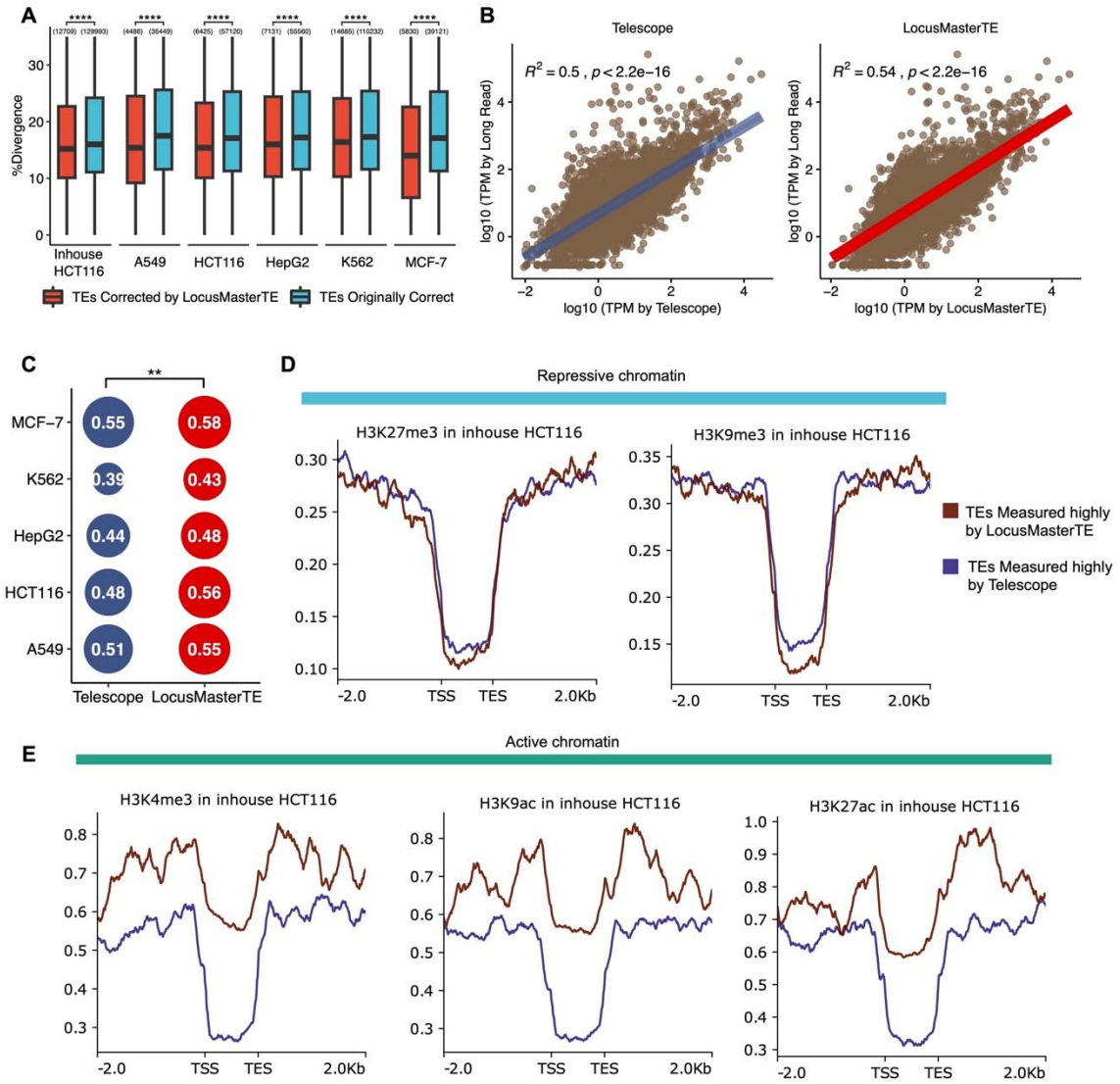

Figure 4: Young and active TEs identified by LocusMasterTE supported by histone markers

(A) Age distribution of TEs across all cell lines between two groups of TEs. Red: TEs quantified differently between LocusMasterTE and Telescope [8] (TEs corrected by LocusMasterTE). Blue: TEs quantified the same across LocusMasterTE and Telescope [8] (TEs initially correct). We conducted a paired t-test for each cell line sample. Quantitative numbers are labeled above each box. (B) Correlation plots for Telescope [8] and LocusMasterTE with long-read inhouse HCT116 cell line. log10 TPM counts are used. (C) Summary bubble plot for correlation with long-read in SG-NEx cell lines. Paired t-test was performed between two groups. (D) Two repressive chromatins fold change over control coverage of in-house HCT116 cell sample for the same two TE groups. Brown: TEs quantified higher in LocusMasterTE (TEs measured highly by LocusMasterTE). Blue: TEs are quantified higher in Telescope [8] (TEs are measured

highly by Telescope). (E) Three active chromatins fold change over control coverage of in-house HCT116 cell line for two TE groups.

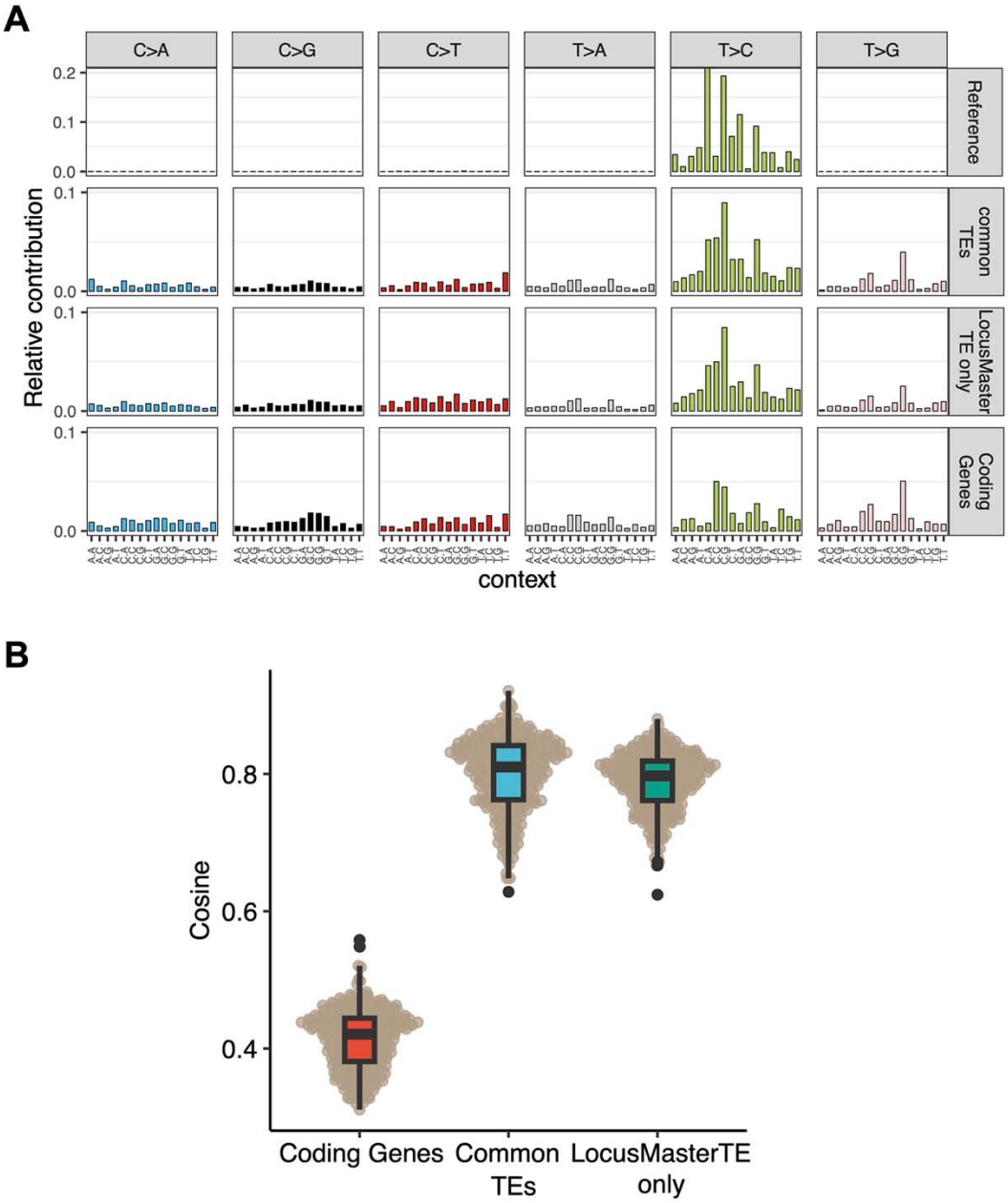

Figure 5: TEs solely captured by LocusMasterTE are active in RNA editing events
(A) Trinucleotide profiles for four groups: ADAR-WT-MDA5-5aza sample with inverted SINEs and ADAR1 [30], TEs commonly detected for both Telescope [8] and LocusMasterTE (common TEs), TEs solely detected by LocusMasterTE (LocusMasterTE only TEs), and coding genes (Coding genes) (B) Cosine similarity values between "common TEs (blue), "LocusMasterTE only TEs (green)", and "coding genes (red)" based on "Reference" for all TCGA-COAD tumor samples.