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

Endogenous transposable elements (TEs) are implicated in human diseases due to their propensity to compromise genome integrity. Although short-read sequencing is now frequently used to examine TE expression, the highly repetitive nature of TEs limits their accurate quantification at the locus-specific level. We have developed LocusMasterTE, an improved method that integrates information from long-read RNA sequencing to enhance TE quantification. The fractional transcript per million (TPM) from long reads serves as a prior distribution during the Expectation-Maximization (EM) model in short-read TE quantification, thereby enabling the reassignment of multi-mapped reads to correct expression values. Using simulated short-reads, our results indicate that LocusMasterTE outperforms existing quantitative approaches and is especially favorable for quantifying evolutionarily younger TEs. Using matched cell line RNA-seq data, we further demonstrate improved locus-specific TE quantification by LocusMasterTE with stronger enrichment in active, and depletion at repressive, histone marks. Finally, by integrating colorectal cancer cell line long-read sequencing data with short read RNA-seq data from The Cancer Genome Atlas colorectal cancer cohort, we demonstrate LocusMasterTE's ability to identify survival-related TEs and uncover new expression associations between locus-specific TEs and neighboring genes. By providing more accurate quantification, LocusMasterTE offers the potential to reveal novel functions of TE transcripts.

1 Introduction

Endogenous transposable elements (TEs) are repetitive genomic units accounting for half of the human genome (Bourque et al. 2018). Based on genetic composition and monophyletic origin, TEs can also be grouped based on genetic composition and monophyletic origin: class > family > subfamily (Bourque et al. 2018). Consensus sequence within the subfamily accounts for the high similarity in sequence compositions of TEs (Jurka and Smith 1988). In normal cells, TEs are firmly regulated through epigenetic control including histone modification, but in cancer cells, an extensive number of TEs are associated with epigenetic dysregulation (Feinberg and Vogelstein 1983). With the accumulation of mutations over time (Piegu et al. 2015), most TEs are now inactive at the

transcriptional level; however, recent studies have identified a few active TEs that have a significant impact on cancer (Slotkin and Martienssen 2007). Moreover, those remaining active TEs influence regulatory elements and have substantial implications in tumorigenesis (Chuong et al. 2017; Vidal et al. 2017).

The importance of TEs has led to the development of tools for their quantification by short RNA-seq data. Some of these include Telescope (Bendall et al. 2019), SalmonTE (Jeong et al. 2018), SQuIRE (Yang et al. 2019), TEtranscripts (Jin et al. 2015), TEtools (Lerat et al. 2017), and RepEnrich (Criscione et al. 2014). However, as TEs are highly repetitive, especially within the same subfamily, there are many cases where one read from short-read RNA sequencing is aligned to numerous TEs with similar sequence compositions (Yang et al. 2019). Especially, newly inserted TEs (i.e. young TEs) generally accumulate less polymorphisms and contain fewer unique sequences leading to high multi-mapping rates (Jin et al. 2015). Some tools avoid the multi-mapping issue by only focusing on TE quantification at the subfamily level. However, considering TE's ability to balance repression and expression, not all TEs in the same subfamily will simultaneously be up or down-regulated (Schwarz et al. 2022). Other tools have utilized the Expectation-Maximization (EM) model to estimate TE expression at specific locus (Lerat et al. 2017). 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 to study genomes and transcriptomes. Due to its longer technical read length, long-read can increase the success rate in TE assembly in the genome, the discovery of novel transcripts from repetitive regions, and the identification of full-length isoforms (Debladis et al. 2017; Shahid and Slotkin 2020).

Given the advantage of long-read over short-read, here, we present LocusMasterTE, long-read assisted short-read TE quantification, a tool that integrates long-read information to quantify TE expression at individual locus levels in short-read RNA sequencing. LocusMasterTE was built upon the Telescope (Bendall et al. 2019) model with the addition of fractional Transcript Per Million (TPM) values as prior terms in the EM model to accurately reassign expression values across TEs. We conducted analyses with LocusMasterTE in three levels: 1) simulated short-read RNA sequencing, 2) cell line samples, and 3) TCGA tissue samples. LocusMasterTE successfully achieves the highest precision and recall in simulated short-read over conventional tools. Further, we conclude that LocusMasterTE correctly captures TEs in relation to histone marks. Generally, TEs exclusively captured by LocusMasterTE are relatively newly inserted elements (i.e. young TEs) and have been demonstrated to be active in RNA editing events in the tri-nucleotide

context. In summary, our study introduces a robust short-read TE quantification tool aided by long-read, which paves a crucial step to understanding TE mechanisms in cancers.

2. Results

2.1. Long-read is superior to short-reads at sequencing young TEs with uniquely mapping reads

Long-read sequencing technique can read fragments typically between 10-100kb even upto >100kb reaching 2Mbp (Jung et al. 2019), thereby greatly decreasing the proportion of multi-mapping during alignment to TE transcriptomes (Fig. 1a). To understand the differences between long-read and short-read for transcriptome sequencing, we performed comparisons at three levels: coding genes, TE subfamily, and TE individuals (detailed in the Methods). As TE subfamily is an aggregation of individual copies, we observed good agreement between long-read and short-read (R^2=0.82); however, there was more substantial difference at locus-specific copy level (R^2=0.41, Fig. 1b). 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. 1).

TEs that have undergone transposition in the genome more recently (i.e. evolutionarily younger TEs) generally have fewer polymorphisms, rendering individual copies more difficult to distinguish and be mapped uniquely. To correctly quantify young TEs using short-reads, diverse TE quantification tools including Telescope (Bendall et al. 2019) and SalmonTE (Jeong et al. 2018) implemented Expectation Maximization (EM) model, which allows multi-mapped reads to be reallocated based on estimated probability. In theory, the EM model should converges with one optimal best hit after rounds of iterations. However, we identified 4% cases where EM ends up with more than two best hits because it has insufficient information about where a read originated from (Fig. 1c). Given the longer read length in long-read sequencing leads to lower multi-mapping rates, we observed that the median age of TEs quantified by long-read was lower than those captured by short-read (158bp/1000bp versus 186bp/1000bp; p-value < 0.0001). The age gap was increased when we compared a group of TEs exclusively quantified by long-read (median = 152) and TEs only counted by short-read (median = 195; p-value < 0.0001) (Fig. 1d). 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).

2.2 LocusMasterTE: Long-read assisted short-read TE quantification tool

Given the benefits of long-read in studying TEs, we devised LocusMasterTE, which integrates long-read information to improve short-read-based TE quantification. As described above, application of EM does not perfectly resolve all multi-mapping reads (Fig.

1c). LocusMasterTE additionally uses long-read information during the EM model, allowing improved resolution of multi-mapping.

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 coordinations of individual TE transcripts 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 transcripts, the read is classified as a multi-mapped read (Fig. 2a). As with Telescope (Bendall et al. 2019), LocusMasterTE represents the proportion of read fragments as a Bayesian mixture model and utilizes the EM model to estimate parameters.

In reference to Telescope (Bendall et al. 2019), 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, inputted 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). To maximize the practically of LocusMasterTE, LocusMasterTE allows users to input any available long-read sample. Although the best accuracy is achievable with matched cell or tissue types between long-read and short-read, considering the case of inputting long-read in a different type of cells or tissues from short-read, LocusMasterTE introduces (--*long_read_weight*), allowing users to manage the weights of long-read information. Lower weights are recommended with distant tissue types, but optimal weights vary between samples. In most cases, inputting between 0.5 and 1.5 for (--*long_read_weight*) is advised. We tested a series of (--*long_read_weight*) values on an inhouse HCT116 short-read cell line sample with a K562 long-read sample from the SG-NEx study (Chen et al. 2021) as input and found that values beyond 0.5 and 1.5 (--*long_read_weight*) had lower accuracy (Supplementary Fig. 2). At the conclusion of the EM optimization, LocusMasterTE provides an updated alignment with finalized counts (Fig. 2d).

2.3. Benmarcking LocusMasterTE on simulated short-read

To demonstrate the accuracy of LocusMasterTE, we simulated short-read RNA-seq datasets based on a long-read RNA-seq sample (Details in Methods). This enabled us to use the long-read data as ground truth. Using simulated short-reads, we compared the accuracy of LocusMasterTE in TE quantification using seven approaches: 1) Unique counts (Fig. 3a), 2) Best Counts (Fig. 3b), 3) RSEM (Li and Dewey 2011) (Fig. 3c), 4)

Telescope (Bendall et al. 2019) (Fig. 3d), 5) SQuIRE (Yang et al. 2019) (Fig. 3e), and 6) SalmonTE (Jeong et al. 2018) (Fig. 3f), 7) LocusMasterTE (Fig. 3g). Both "unique counts" and "best counts" were provided by Telescope (Bendall et al. 2019); the difference mainly lies in handling multi-mapping reads. "Unique counts" discard all multi-mapping reads; whereas, "best counts" randomly assign multi-mapping reads to the highest score during alignment. All TEs listed in RepeatMasker (Chen 2004) were inputted except for simple repeats. For fair comparisons, salmonTE (Jeong et al. 2018) was modified to quantify TEs at individual levels, as it was originally designed to calculate TE at subfamily.

Prior to comparing TE counts, we tested the quality of simulated short-read using coding genes (Supplementary Fig. 3). Out of 19120, 18465 coding genes were consistent between long-read and simulated short-read (TP = 13229, TN = 5236, FP = 120, FN = 535). LocusMasterTE was shown to have 0 falsely detected TEs with highest number of correctly detected TEs (47217), demonstrating highly precise TE quantification among seven approaches (Precision: 1.0, Recall: 0.908; Fig. 3h). SQuIRE (Yang et al. 2019) captures similar amount of TEs with LocusMasterTE, but it overcaptures false positive TEs, leading to low precision (Precision: 0.49, Fig. 3h). As LocusMasterTE was built upon Telescope (Bendall et al. 2019), we carefully compared the performance of the two tools. LocusMasterTE achieved much higher TP and TN with lower FP and FN values, outperforming Telescope (Bendall et al. 2019) (Fig. 3d,g). With simulated short-read samples, LocusMasterTE is demonstrated to be the most accurate among the seven approaches.

2.4. TEs uniquely quantified by LocusMasterTE are evolutionarily younger and marked by active chromatin

To understand the distinctive characteristics of LocusMasterTE, we used cell line samples with matched long and short-read RNA sequencing data as input for LocusMasterTE. This included one inhouse HCT116 sample and five cancer cell lines (A549, HCT116, HepG2, K562, and MCF-7) from the SG-Nex (Chen et al. 2021). First, we examined TEs that showed expression changes between LocusMasterTE and Telescope (Bendall et al. 2019) and found that these were generally evolutionarily younger TEs compared with those that remain unchanged by both methods. This suggests that LocusMasterTE is indeed correcting the expression of difficult to quantify young TEs (Fig. 4a). TE counts generated by Telescope (Bendall et al. 2019) and LocusMasterTE were then compared with long-read results to confirm successful integration of long-read in cell line samples. We observed increased correlation (R^2 = 0.39) with long-read results using LocusMasterTE compared with Telescope (Bendall et al. 2019) (R^2 = 0.37) (Fig. 4b). This trend was observable across all cell types (Fig. 4e, supplementary Fig. 4A).

Histone modifications are markers for TEs activity. We grouped TEs into two, TEs with higher expression by LocusMasterTE and Telescope (Bendall et al. 2019), respectively

(Details in Methods), then compared their relationships with five histone marks, composed of three active (H3K4me3, H3K9ac, H3K27ac) and two repressive (H3K27me3 and H3K9me3). Between two groups of TEs, TEs highly captured by LocusMasterTE had more enrichment for all three active marks (Fig. 4c) and less enrichment for the two repressive marks (Fig. 4d). The same trends was observed across all cell lines (Fig. $4\mathrm{f}, 4\mathrm{g}$, supplementary Fig. 4B). This suggests that LocusMasterTE improves TE qualification as it is more consistent with independent epigenetic data.

2.5. Biological Connection to tissue samples using TCGA

To highlight the biological importance of using LocusMasterTE, we quantified TEs using LocusMasterTE in TCGA-COAD samples (Tomczak et al. 2015), examining the relationship of quantified TEs with: 1) gene enrichment in DNA/RNA sensing genes, 2) correlation with coding genes, 3) RNA-editing, and 4) patient prognosis.

As TEs are frequently related to the immune response in human (Zhu et al. 2021), we performed gene set variation analysis (GSVA) (Hanzelmann et al. 2013) on 57 DNA and RNA sensing genes that are associated with the innate immune system (Supplementary Table 1). Two TE expression matrices generated by Telescope (Bendall et al. 2019) and LocusMasterTE were correlated using GSVA scores of the DNA and RNA sensing pathway (Details in Methods). With two distinctive TE matrices generated by Telescope (Bendall et al. 2019) and LocusMasterTE and 57 DNA/RNA sensing pathway GSVA score matrix, quantified counts by LocusMasterTE (mean R^2 = 0.6) showed a higher correlation with GSVA scores of DNA/RNA sensing pathway than by Telescope (Bendall et al. 2019) (mean R^2 = 0.57) (Fig. 5a).

As TEs impose great influences on nearby genes, especially in cancer, we performed a series of correlations between TEs and nearby, $+/-5kb$, coding genes. To filter only autonomous TE expression from host genes or intron retention (Smart et al. 2018), intergenic TEs were further used in correlation analysis. For each set of TEs and genes, TEs were labeled based on the relative position of TEs with genes: upstream and downstream. Similar to GSVA analysis, quantified TEs using Telescope (Bendall et al. 2019) and LocusMasterTE were used and resulted correlation values (Kumar et al. 2018) were compared. For both upstream and downstream TEs, LocusMasterTE discovered important TE and gene relationships (|spearman correlation estimates(ρ) |>= 0.5) more frequently than Telescope (Bendall et al. 2019) (Fig. 5b, 5c). A specific example was drawn to study the difference between TEs detected by Telescope (Bendall et al. 2019) and LocusMasterTE. Prior to direct comparisons, we checked whether TEs were freed from intron retention or influenced by nearby genes. Specific copies of L1ME3A (L1ME3A; chr1:54,639,446-54,639,578) did have distinctive reads (Fig. 5d). Using quantified values by LocusMasterTE of L1ME3A; chr1: 54,639,446-54,639,578 showed to have statistically important correlation ($\rho = 0.64$) with MROH7 (ENSG00000184313) (Fig. 5e). In

contrast, using counts generated by Telescope (Bendall et al. 2019), the correlation between L1ME3A; chr1: 54,639,446-54,639,578 and MROH7 was not higher than 0.5 ($\rho = 0.36$) (Fig. 5f).

RNA editing refers to the post-transcriptional modification of RNA sequences in a few positions. It frequently occurs in specific contexts (Wang et al. 2018). Among diverse families of TEs, we focused on RNA editing on Alu elements at tri-nucleotide. We used reference RNA editing signatures derived from MDA5 protected Alu elements from 5-azacytidine treated ADAR wild-type samples (Mehdipour et al. 2020) and compared with three groups: 1) coding genes, 2) TEs commonly detected by both Telescope (Bendall et al. 2019) and LocusMasterTE, labeled as “common”, and 3) TEs exclusively captured by LocusMasterTE, labeled as “only in LocusMasterTE”. The mutational profile of TCGA-COAD samples demonstrates high similarity among reference editing sites, “common” and "only in LocusMasterTE” groups by having peaks at CTA $>$ CCA, CTG $>$ CCG, and GTG $>$ GCG (Fig. 5g). Moreover, cosine similarity showed high similarity in expected RNA editing signatures, “common” and "only in LocusMasterTE” with a clear difference in coding genes (Fig. 5h).

Finally, we utilized LocusMasterTE to observe the connection in the survival data of TCGA samples. To highlight the advantage of LocusMasterTE, we selected TEs which were differently quantified between LocusMasterTE and Telescope (Bendall et al. 2019) (Details in Methods). Moreover, among all TEs, only intergenic or intronic TEs were used. Survival analysis using lasso-cox regression models was performed on a progression-free interval (PFI) (Tibshirani 1997). After 100 iterations, we filtered out the two most commonly occurring TEs (Supplementary Table 2). Based on expressions of 5 TEs, patients were grouped into two following the greatest delta change under the CDF curve provided by ConsensusClusterPlus (Wilkerson and Hayes 2010) (Supplementary Fig. 5A). The numbers of two groups of patients were also balanced (Supplementary Fig. 5B). Based on expressions of two TEs, patients were grouped into two clusters. There was higher LSU-rRNA copy expression in group 1 compared with group 2 but lower THE1C TE copy expression (Fig. 5I). We found out two groups separated by two TEs have different prognosis in PFI index (log-rank p-value = 0.017) (Fig. 5J). To examine independent expression from nearby or derived genes, we looked at the correlation in pairs of TEs and their located genes. Both two TEs demonstrate distinctive expressions (Supplementary Fig. 5C). We further conducted the same survival analysis for those genes but could not get a significant log-rank result (log-rank p-value = 0.92), highlighting independent expressions of 2 TEs (Supplementary Fig. 5D-E). In the two survival-related TEs, THE1C; chr2: 16797244-16797271 was exclusively captured by LocusMasterTE (Supplementary Fig. 5F-G), suggested that TE expressions by LocusMasterTE can better connect to clinical outcomes.

3. Discussion

Transposable elements impose essential roles in cancer and complex biological diseases in humans (Bourque et al. 2018). However, because TEs have repetitive genomic sequences, it has been challenging to align and quantify TEs accurately. Since long-read has an advantage over short-read in technical read length, we developed LocusMasterTE to integrate the long-read results during EM. The inclusion of long-read results fosters the EM model to align reads more correctly and faster than conventional short-read TE quantification tools. Especially, LocusMasterTE overcomes cases where EM concludes with more than two best hits by providing long-read information additionally. Also, LocusMasterTE allows users to use any long-read regardless of cell or tissue type. This feature further makes LocusMasterTE the best practice.

The accuracy of LocusMasterTE was first proved by simulated short-read. Among diverse TE quantification approaches, LocusMasterTE successfully achieves 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 (Bendall et al. 2019) , TEs quantified by LocusMasterTE were generally younger. TEs quantified by LocusMasterTE was also more consistent with active and repressive histone marks quantified idepenedently by ChIP-seq (Consortium et al. 2020). We also find the biological importance of utilizing LocusMasterTE in the TE study. We successfully identified two prognostic TEs quantified by LocusMasterTE. Moreover, LocusMasterTE has revealed more statistically significant relationships between intergenic TEs and genes. Analyses on cell lines and tissue samples both demonstrate precise TE quantification of LocusMasterTE.

Finally, while we believe LocusMasterTE is robust and accurate for TE quantification, it still holds a few limitations. As LocusMasterTE is a long-read aided tool, long-read information is mandatory. However, considering the high cost of long-read sequencing compared to other sequencing techniques, not all samples have both short-read and long-read simultaneously, or at least where long-reads are sequenced to sufficient depth. Although LocusMasterTE allows users to input any type of long-read samples, the best performance is still achievable when inputting long-read from the same sample. In the case of inputting the same reference long-read sample and running LocusMasterTE in different short-read samples, TE counts by LocusMasterTE may hold few estimated counts. Secondly, the performance of LocusMasterTE depends on the coverage in long-read. With deeper depth, LocusMasterTE quantifies TEs more accurately. Moreover, although the optimal range for (--*long_read_weight*) is 0.5 to 1.5, the performance of this parameter depends on long-read sequencing depth. In spite of these limitations, LocusMasterTE

achieves the best TE quantifications among available tools, paving the way for discovering unknown human TE functions.

4. Methods

4.1. Adding Long-read information as prior distribution in EM model

In this study, LocusMasterTE was developed based on the fragment reassignment mixture model, Telescope (Bendall et al. 2019), to enable locus-specific quantification of transposable elements. Specifically, LocusMasterTE utilizes an expectation maximization (EM) algorithm (Levitan and Herman 1987) 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 (Li 2018). Following alignment, TE quantification of long-read is performed using featureCounts (Liao et al. 2014) with a GTF file, such as one generated by RepeatMasker (Chen 2004), containing individual TE transcripts $T = \{t_1, t_2, \dots, t_K\}$, where K is the number of transcripts. Raw counts for each TE transcript generated by featureCounts (Liao et al. 2014) 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 (Bendall et al. 2019), $\pi = \{\pi_1, \pi_2, \dots, \pi_K\}$ and $\theta = \{\theta_1, \theta_2, \dots, \theta_K\}$ are two sets of parameters: representing the proportion of observed fragments origining from each transcript, and a similar reasignment parameters for multi-mapped reads, respectively (Schwarz et al. 2022). 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 estimiated 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 as inputted value for (*--long_read_weight*) option.

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

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

From the mixture model provided by Telescope (Bendall et al. 2019), introducing a new parameter T^{Lf} during the EM model changes the parameters π and θ, which consequently affects overall mixture weights in the model (Bendall et al. 2019). 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 $\epsilon < 0.001$ (Wu 1983). All other formulas are in accordance with the Telescope (Bendall et al. 2019).

4.2. Other essential features in LocusMasterTE

1. Due to differences in sequencing techniques, certain TEs are only captured by either short-read sequencing or long-read sequencing. However, since LocusMasterTE directly utilizes long-read TPM counts and updates parameters through multiplication, when quantified TPM counts from long-read are 0, no matter how much TEs are captured by short-read, final counts will become 0. To rescue TEs only quantified by short-read sequencing, (*--rescue_short*) argument is introduced to replace 0 with small constant in long-read TPM count. The default value sets to 0, but users can freely input any small constant (i.e. $1e^{-50}$) based on their preferences.

2. To resolve cases where there are multiple potential best reassignments, even after completion of EM model, as a result of the (*--reassign_mode*) argument originally introduced by Telescope (Bendall et al. 2019), we have added an extra option named "long_read". Similar to the average choice given by Telescope (Bendall et al. 2019), 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.

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

4.3. Generation of HCT 116 RNA sequencing data

HCT116 cells were cultured in McCoy's 5A media supplemented with 10% FBS and seeded into 35 cm dishes. At 80% confluence, RNA was extracted using Qiagen RNeasy Mini kit (74104) following manufacturer's directions. For short-read RNA sequencing, samples were sent to Novogene Co., Ltd (Beijing) for library preparation and illumina sequencing. Briefly, mRNA was enriched with oligo (dT) beads and then fragmented randomly. cDNA was synthesised from fragmented mRNA using random hexamers as primers and then subject to second-strand synthesis, end-pair, dA-tailing and ligation of sequencing adapters. Sequencing was performed on an Illumina Novaseq 6000 instrument.

For long-read RNA sequencing, 50 ng was used for library preparation using Oxford Nanopore cDNA-PCR barcoding kit (Oxford Nanopore, SQK-PCB109), performed according to manufacturer's protocol assuming an average transcript length of 1.5 Kb. 100 fmol of library was loaded on 1 flow cell (Oxford Nanopore, FLO-MINSP6) which was sequenced on the MinION Sequencing Device (Oxford Nanopore, MIN-101B).

4.4. Information about TEs

TEs can be divided into three categories based on genomic locations: exon, intron, and intergenic. As some TEs overlap multiple regions, locations were prioritized by exon, intron, and intergenic. As some TEs are overlapping between regions, locations were prioritized by exon, intron followed by intergenic. BEDTools (Quinlan and Hall 2010) with Gencode annotation of hg38 version 41 Basic GTF file (Harrow et al. 2012) was used to extract chromosome locations for exon, intron, and intergenic, and each TEs were labeled based on their genomic locations.

Using the subfamily information provided by RepeatMasker, aggregation to subfamily level was done by grouping TEs individuals based on annotated subfamily information (Chen 2004). Moreover, TEs in simple repeat class were filtered out for all analyses, due to their insignificance.

Throughout our study, we conducted age analysis several times. For all cases, age information was provided by the RepeatMasker database in the "milliDiv" column which represents a number of bases mismatches in parts per thousand (Chen 2004).

4.5. Gene and TE quantification

For all analyses where genes were involved, we only selected coding genes. We obtained the coding gene GTF file from the UCSC table browser by applying MANE annotated gene filters (Morales et al. 2022). GTF file for TE quantification had been generated using RepeatMasker for Hg38 mentioned above (Chen 2004). Minimap2 (Li 2018) with Genome Reference Consortium Human Build 38 (GRCh38) (Schneider et al. 2017) fasta were used for aligning long-read. For short-read alignment, we used a STAR (Dobin et al. 2013) as

an aligner with suitable parameters provided in GitHub. With generated BAM file, TE quantification was done using LocusMasterTE for all the analyses, except for long-read and short-read comparison in Figure 1. Having a fair comparison between long-read and short-read, we utilized featureCounts to assign genomic features to reads for both long-read and short-read (Liao et al. 2014). Moreover, because LocusMasterTE originated from Telescope, many comparisons were done between Telescope and LocusMasterTE (Bendall et al. 2019).

In most of the cases, instead of raw counts, TPM (Transcripts Per Kilobase Million) was used throughout the study. For TPM library size, the sum of coding genes were used.

4.6. BWA

For short-read, STAR aligner was first used to align against the whole genome (Dobin et al. 2013). Using bedtools, regions with basic genes have been filtered out (Quinlan and Hall 2010). The remaining reads were remapped with BWA (Li et al. 2009), and regions overlapping with TEs were selected. In case of long-read, minimap2 was used for initial alignment (Li 2018). With unmapped reads, BWA was remapped and read for TEs were extracted. The number of multi-mapping reads and uniquely mapped reads were calculated based on the "FLAG" tag in SAM (Cock et al. 2010).

4.7. Downsampling of Short-Read

As short-read RNA sequencing data has greater coverage depth compared with long-read, to conduct an unbiased comparison, we down-sampled short-reads by matching the number of mapped bases. SAMtools was first used to calculate the number of bases mapped and again used to generate subsample reads from short-reads (Cock et al. 2010). As described above, gene and TE quantification were performed and compared between long-read and short-read.

4.8. Simulated short-read RNA sequencing

For benchmarking LocusMasterTE, a simulated short-read was generated from long-read RNA sequencing using the read simulator Spanki (Sturgill et al. 2013). Inhouse HCT 116 long-read sample was used as input for 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 long-read TPM values for each TE as coverage (Sturgill et al. 2013).To match the number of bases in generated FASTQ (Cock et al. 2010) file for simulated short-read and long-read, we first got 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 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.

4.9. Comparison with other tools

To compare the accuracy of LocusMasterTE, TEs in simulated short-read RNA sequencing were quantified using Telescope (Bendall et al. 2019), RSEM (Li and Dewey 2011), SQuIRE (Yang et al. 2019), and SalmonTE (Jeong et al. 2018). For a fair comparison, the same GTF file composed of TE individuals described above was used for all tools. To generate an index for SalmonTE, fasta file consisting of TE individual sequences, which was generated by a locus-specific TE expression study, was inputted (Schwarz et al. 2022). Same as the aligner for LocusMasterTE. STAR was used for alignment in Telescope (Bendall et al. 2019) and SQuIRE (Yang et al. 2019). For RSEM (Li and Dewey 2011), bowtie2 (Langmead and Salzberg 2012) was used as an aligner. For comparisons, all values were converted into log (TPM) with the addition of $1e^{-7}$ to prevent an infinite number while logging. Except for the tools where TPM values were provided such as SalmonTE (Jeong et al. 2018), the sum of coding genes counts was used as library size.

As a simulated short-read was generated from a long-read RNA sequencing sample, quantified TE counts from a long-read were considered as ground truth. To remove TEs unsuccessfully simulated, TEs captured by at least one approach were only considered. Compared with ground truth, based on the confusion matrix generated for each tool, accuracy and recall values were calculated.

4.10. Cell line analysis

In addition to inhouse HCT 116, data from SG-nex study (Chen et al. 2021) was also used in cell line analysis. In SG-nex study, there are 5 different cell line samples with both short-read and long-read RNA sequencing available (Chen et al. 2021). For cell line analysis to TCGA analysis, parameters for LocusMasterTE were set as below.

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

As LocusMasterTE is grounded in Telescope (Bendall et al. 2019), all the comparisons were conducted with Telescope (Bendall et al. 2019). For each TE, there were two values generated: counts from Telescope (Bendall et al. 2019) ($TE^{c.tel}$) and counts from LocusMasterTE $\left(TE^{\text{c.las}}\right)$.

4.10.1. Age analysis

In age analysis, TEs were divided into two groups. TEs having the same counts between Telescope (Bendall et al. 2019) and LocusMasterTE were considered as "originally correctly measured TEs". For TEs quantified differently between Telescope (Bendall et al. 2019) and LocusMasterTE were defined as "corrected by LocusMasterTE", because those TEs were either over-estimated or under-estimated in Telescope (Bendall et al. 2019).

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

T-test was conducted between two groups.

4.10.2. Histone analysis

Five histone modifications representing active (H3K4me3, H3K27ac, and H3K9ac) and inactive (H3K27me3 and H3K9me1) chromatin were used for histone analysis. BigWig files from ChIPseq data of the 5 histone marks for each cell line were downloaded from ENCODE project (Consortium et al. 2020). In histone analysis, TEs were re-divided into two groups.

- $TE^{\text{c.las}} > TE^{\text{c.tel}}$: TEs having higher expression in LocusMasterTE
- $TE^{\text{c.las}} < TE^{\text{c.tel}}$: TEs having higher expression in Telescope (Bendall et al. 2019)

For better comparison, TEs with non-trivial differences between Telescope (Bendall et al. 2019) and LocusMasterTE $\left|TE^{\text{c.las}} - TE^{\text{c.tel}}\right| > 1$ were selected. Then, we used computeMatrix from deepTools (Ramirez et al. 2014) to calculate scores.

4.11. TCGA samples

We utilized TCGA-COAD study samples, composed of 41 normal samples and 310 tumor samples, a total of 351 samples from 305 different patients (Tomczak et al. 2015). As all analyses requires tumor samples, only tumor samples were used for other analyses. As TCGA-COAD samples (Tomczak et al. 2015) only have short-read RNA sequencing, we inputted long-read from HCT 116 sample as long-read input for LocusMasterTE. Same as cell line analysis, all the comparisons were performed between LocusMasterTE and Telescope (Bendall et al. 2019).

4.11.1. GSVA score for DNA/RNA Sensing gene

The list of DNA/RNA sensing genes is composed of cytosolic DNA-sensing pathway for hg38 in the KEGG database (Kanehisa and Goto 2000) with the addition of EIF2AK2 (i.e. PKR) (Bullido et al. 2008) and IFIH1 (i.e. MDA5) (Triantafilou et al. 2012), which are proteins well known to recognize and bind dsRNA; thus a total of 57 genes were included in the list (Supplementary Table 1).To compute the enrichment score for the DNA/RNA sensing pathway, Gene Set Variation Analysis (GSVA) score was calculated for all genes in the list (Hanzelmann et al. 2013). We inputted Illumina gene expression RNAseq from the UCSC Xena database to calculate GSVA scores (Hanzelmann et al. 2013). With GSVA scores of 57 genes for TCGA-COAD patients, we performed Spearman's (Kumar et al. 2018) correlation between GSVA scores and TEs counts quantified by Telescope (Bendall et al. 2019) and LocusMasterTE. TPM values for TEs were generated from raw counts for both tools. Prior to conducting correlation, two filters were applied to the TEs expression matrix: 1) remove rows with more than 30% to be 0 and 2) remove rows with low variance (90% of samples having the same value for one TE). Since there can be both positive and negative correlations between genes and TEs, for each DNA-RNA sensing gene, TEs with top 10 and low 10 correlation values were selected. This procedure was performed on TE counts produced by Telescope (Bendall et al. 2019) and LocusMasterTE , and two correlation values were compared.

4.11.2. Correlation between Genes and TEs

To conduct a fair correlation between genes and TEs, Illumina gene expression RNA sequencing from UCSC Xena database was first converted into TPM values (Karolchik et al. 2003). Coding genes were only selected using MANE annotated gene filters from UCSC as explained in Gene and TE quantification section (Morales et al. 2022). To compare the performance between Telescope (Bendall et al. 2019) and LocusMasterTE, we used two TE matrices quantified by two tools and converted raw final counts into TPM values. We specifically chose intergenic TEs to find important correlations between coding genes. Spearman's correlation (Kumar et al. 2018) was performed between coding genes that lie within $5\ \mathrm{kb} \pm$ from the gene body of intergenic TEs, determined by BEDTools (Quinlan and Hall 2010). With the resulting correlation values, locations of intergenic TEs were grouped as upstream or downstream based on the relative position between intergenic TEs and correlated genes. The whole procedure was applied for both LocusMasterTE and Telescope (Bendall et al. 2019) and the difference between correlation values for the two tools was evaluated using the Chi-square test (Ugoni and Walker 1995). Let TEs having higher correlation values in LocusMasterTE as $TE_{\mathrm{cor}}^{\mathrm{las}}$ and TEs with higher correlation values in Telescope (Bendall et al. 2019) as TE_{cor}^{tel}, and 2×2 table was generated for four groups: $\left|TE_{\mathrm{cor}}^{las} < 0.5\right|, \left|TE_{\mathrm{cor}}^{las} \geq 0.5\right|, \left|TE_{\mathrm{cor}}^{tel} < 0.5\right|$, and $\left|TE_{\mathrm{cor}}^{tel} \geq 0.5\right|$. Finally, the chi-square test was applied in a 2×2 table (Ugoni and Walker 1995).

4.11.3. RNA Editing analysis

TEs were first grouped into two and their genomic locations were extracted for individual TCGA-COAD samples.

1 $TE^{\text{c.las}} \neq 0 \wedge TE^{\text{c.tel}} = 0$: TEs exclusively measured by LocusMasterTE

2 $TE^{\text{c.las}} \neq 0 \wedge TE^{\text{c.tel}} \neq 0$: TEs both measured by two tools

35000 TEs were randomly selected for both groups. Additionally, 35000 coding genes were randomly drawn for comparison. For a clear comparison, we only extracted exons of coding genes and filtered out UTRs and introns. Moreover, we used the ADAR-WT-MDA5-5aza sample provided by an epigenetic therapy study in inverted SINEs and ADAR1 (Mehdipour et al. 2020) to extract expected editing signatures, and used this sample as "reference". With extracted genomic location files and GRCh38 fasta same as above mentioned (Schneider et al. 2017), we used bam-readcount to generate information at specific nucleotide positions (Khanna et al. 2021). Germline SNPs overlapping the gnomAD database (Gudmundsson et al. 2022) were filtered out using bcftools (Narasimhan et al. 2016). Except for the coding genes group, we only used regions for Alu in all groups. MutationalPatterns from the R package was used to draw mutational profiles for all groups in trinucleotide contexts (Blokzijl et al. 2018). We further performed a series of cosine similarity tests using lsa R package (Evangelopoulos 2013) between all groups and compared for whole TCGA-COAD samples (Chou and Hsu 2018).

4.11.4. Survival analysis

Prior to survival analysis, to highlight the difference between LocusMasterTE and Telescope (Bendall et al. 2019) only TEs differently captured by two tools were filtered, $\left|TE^{\text{c.las}} - TE^{\text{c.tel}}\right| > 0.1$. Moreover, intronic and intergenic TEs were considered. The survival database on PFI in COAD patients was obtained by UCSC Xena (Harrow et al. 2012). Prior to survival analysis, only TEs expressed more than 80% of TCGA- COAD samples were only selected. Based on selected TEs, we applied the lasso cox regression model (Tibshirani 1997). The model ran 100 times and we selected the most selected TEs with FDR < 0.05 and p-value < 0.05, and finally ended up with 2 TEs. The optimal number of clusters was determined using ConsensusClusterPlus *R* packages (Wilkerson and Hayes 2010), and by using the survival R package, the survival plot was drawn.

5. Data and Code Availability

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 are retrieved from ENCODE [https://www.encodeproject.org/]. Short-read and Long-read 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 GSE225380 and GSE225377 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 Interest Statement

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

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. 2019. Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol* **15**: e1006453.

Blokzijl F, Janssen R, van Boxtel R, Cuppen E. 2018. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**: 33.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvak Z, Levin HL, Macfarlan TS et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**: 199.

Bullido MJ, Martinez-Garcia A, Tenorio R, Sastre I, Munoz DG, Frank A, Valdivieso F. 2008. Double stranded RNA activated EIF2 alpha kinase (EIF2AK2; PKR) is associated with Alzheimer's disease. *Neurobiol Aging* **29**: 1160-1166.

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

Chen Y, Davidson NM, Wan YK, Patel H, Yao F, Low HM, Hendra C, Watten L, Sim A, Sawyer C et al. 2021. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv* doi:10.1101/2021.04.21.440736: 2021.2004.2021.440736.

Chou EP, Hsu SM. 2018. Cosine similarity as a sample size-free measure to quantify phase clustering within a single neurophysiological signal. *J Neurosci Methods* **295**: 111-120.

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

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**: 1767-1771.

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

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

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

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Evangelopoulos NE. 2013. Latent semantic analysis. *Wiley Interdiscip Rev Cogn Sci* **4**: 683-692.

Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**: 89-92.

Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, Genome Aggregation Database C, Rehm HL, MacArthur DG, O'Donnell-Luria A. 2022. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**: 1012-1030.

Hanzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**: 7.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

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

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

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

Jurka J, Smith T. 1988. A fundamental division in the Alu family of repeated sequences. *Proc Natl Acad Sci U S A* **85**: 4775-4778.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27-30.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**: 51-54.

Khanna A, Larson DE, Srivatsan SN, Mosior M, Abbott TE, Kiwala S, Ley TJ, Duncavage EJ, Walter MJ, Walker JR et al. 2021. Bam-readcount -- rapid generation of basepair-resolution sequence metrics. *ArXiv*.

Kumar N, Kumar P, Badagabettu SN, Lewis MG, Adiga M, Padur AA. 2018. Determination of Spearman Correlation Coefficient (r) to Evaluate the Linear Association of Dermal Collagen and Elastic Fibers in the Perspectives of Skin Injury. *Dermatol Res Pract* **2018**: 4512840.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

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

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

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

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

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

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

Mehdipour P, Marhon SA, Ettayebi I, Chakravarthy A, Hosseini A, Wang Y, de Castro FA, Loo Yau H, Ishak C, Abelson S et al. 2020. Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. *Nature* **588**: 169-173.

Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva O, Farrell CM et al. 2022. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**: 310-315.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**: 1749-1751.

Piegu B, Bire S, Arensburger P, Bigot Y. 2015. 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* **86**: 90-109.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

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

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849-864.

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

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

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272-285.

Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, Fugmann T, Wong KK, Van Allen EM. 2018. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* **36**: 1056-1058.

Sturgill D, Malone JH, Sun X, Smith HE, Rabinow L, Samson ML, Oliver B. 2013. Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics* **14**: 320.

Tibshirani R. 1997. The lasso method for variable selection in the Cox model. *Stat Med* **16**: 385-395.

Tomczak K, Czerwinska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**: A68-77.

Triantafilou K, Vakakis E, Kar S, Richer E, Evans GL, Triantafilou M. 2012. Visualisation of direct interaction of MDA5 and the dsRNA replicative intermediate form of positive strand RNA viruses. *J Cell Sci* **125**: 4761-4769.

Ugoni A, Walker BF. 1995. The Chi square test: an introduction. *COMSIG Rev* **4**: 61-64.

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

Wang Y, Park S, Beal PA. 2018. Selective Recognition of RNA Substrates by ADAR Deaminase Domains. *Biochemistry* **57**: 1640-1651.

Wilkerson MD, Hayes DN. 2010. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**: 1572-1573.

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

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

Zhu X, Fang H, Gladysz K, Barbour JA, Wong JWH. 2021. Overexpression of transposable elements is associated with immune evasion and poor outcome in colorectal cancer. *Eur J Cancer* **157**: 94-107.

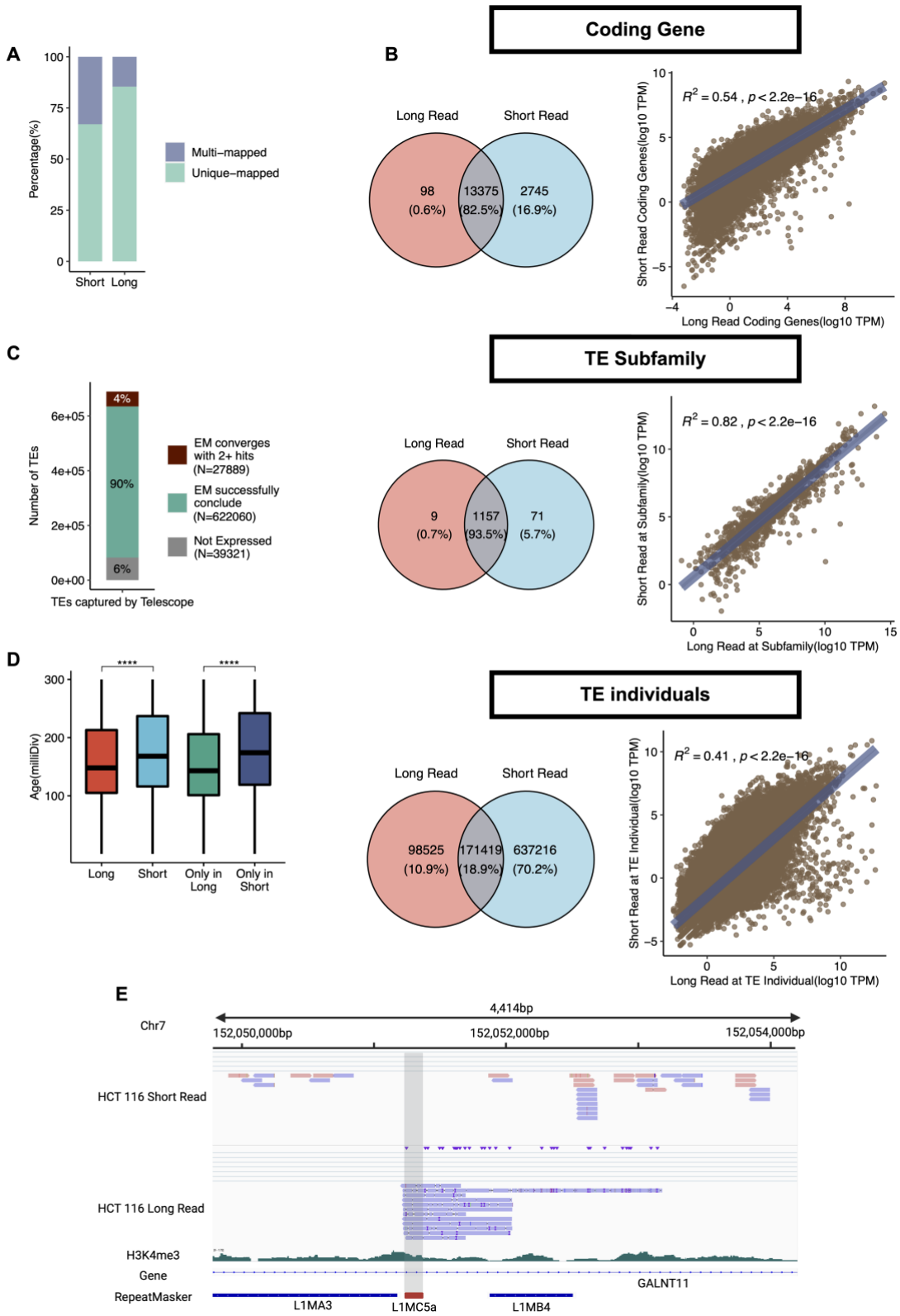

Figure 1: Comparison between long-read and short-read sequencing techniques

(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 three domains: coding genes, TE subfamily, and TE individuals. Right: Correlation plots between log10 TPM count for long-read and short-read in three same domains. (C) Overview of cases where EM takes part in short-read TE quantification by Telescope (Bendall et al. 2019). The total number of TEs corresponds to TEs initially aligned by Telescope (Bendall et al. 2019). 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 originated reads. (D) 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 exclusively captured in long-read (Only in Long). Dark blue: TEs 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.

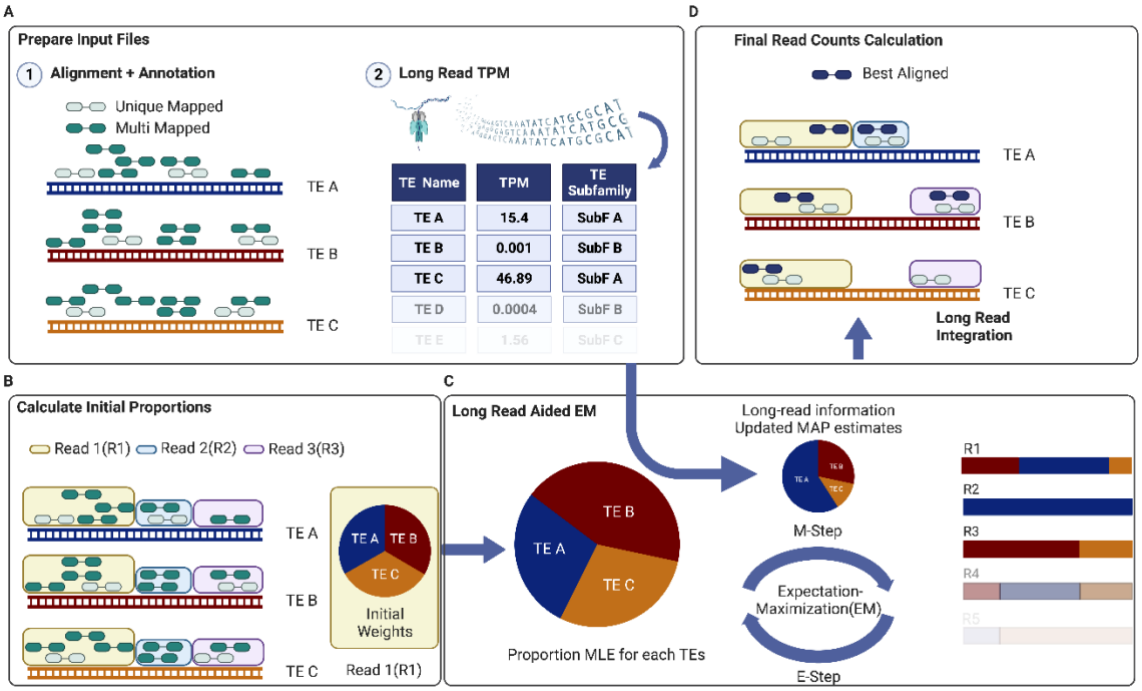

Figure 2: Schematic overview of LocusMasterTE

An example illustration for LocusMasterTE pipeline. (A) Short-read RNA sequencing is aligned using provided annotation file. Based on the alignment score, reads are classified into uniquely mapped reads and multi-mapped reads. For a quick demonstration, uniquely mapped reads and multimapped reads are aligned to TE A, TE B, and TE C in this figure. In the meantime, long-read RNA sequencing is aligned by minimap2 (Li 2018) and TEs are quantified using FeatureCounts (Liao et al. 2014) with the same annotation file provided for short-read RNA sequencing. TPM counts for long-read are calculated, and belonging subfamily are provided for each TE individual. (B) Initially, equal weights are assigned. For instance, as Read 1 (R1) is aligned across all TE A, TE B, and TE C, and the same weights are initialized for all three TEs (TE A: 0.33, TE B: 0.33, TE C: 0.33). If one read is uniquely assigned to one TE, then we assign 1 to a uniquely mapped TE. (C) Based on the initialized weights, the EM model runs until it converges. During the E-step, we calculate MLE for all TEs. In M-step, MAP estimates are updated with the integration of long-read information calculated in (A). (D) Finally after EM model coverages, best-aligned read counts are determined leading to the final assignment of TE counts.

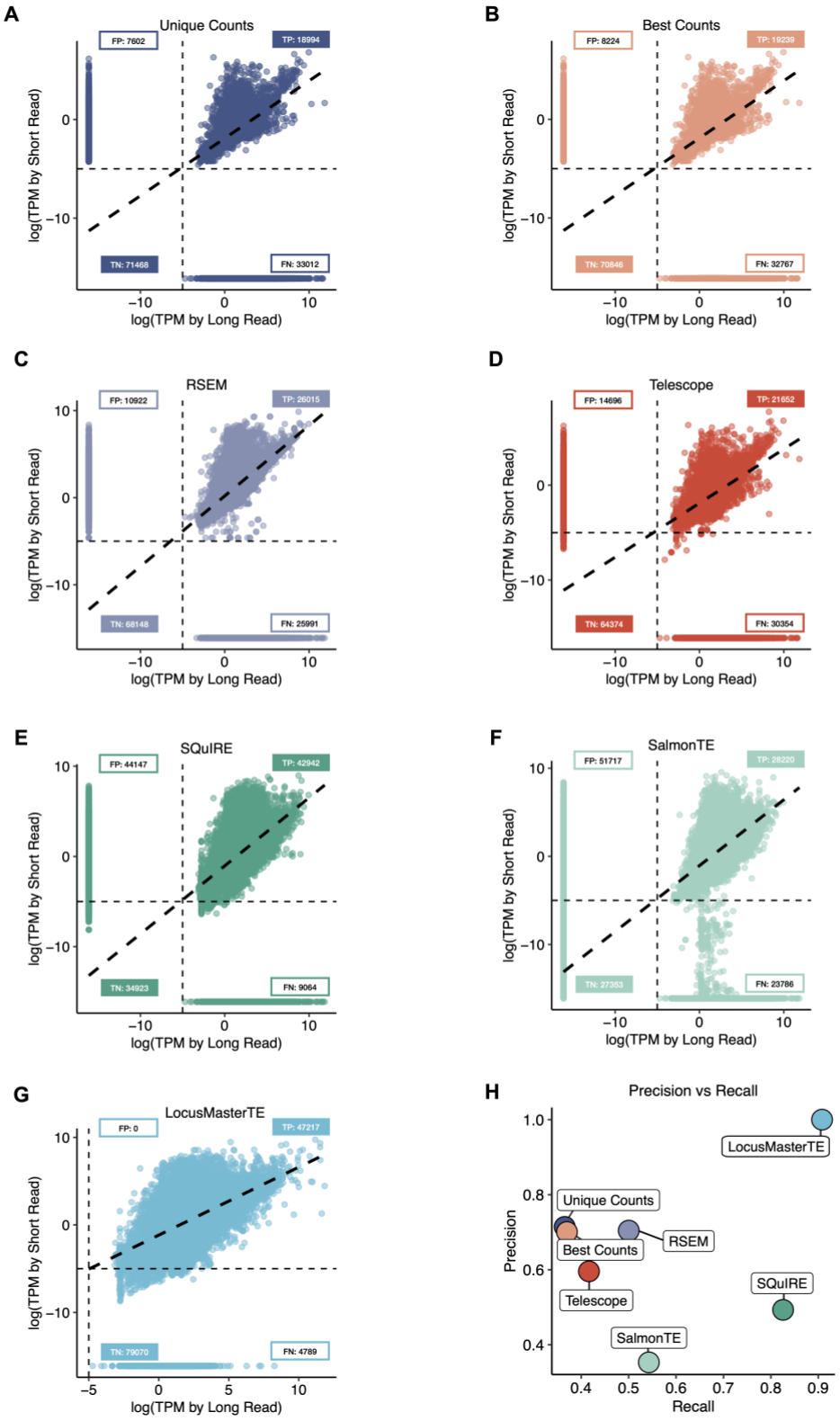

Figure 3: Assessing the accuracy of LocusMasterTE in simulated short-read

7 distinctive approaches to quantify TEs in a simulated short-read. Tested approaches are (A) unique counts provided by Telescope (Bendall et al. 2019), (B) best counts provided by Telescope (Bendall et al. 2019), (C) RSEM (Li and Dewey 2011), (D) Telescope (Bendall et al. 2019), (E) SQuIRE (Yang et al. 2019), (F) salmonTE (Jeong et al. 2018), 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.

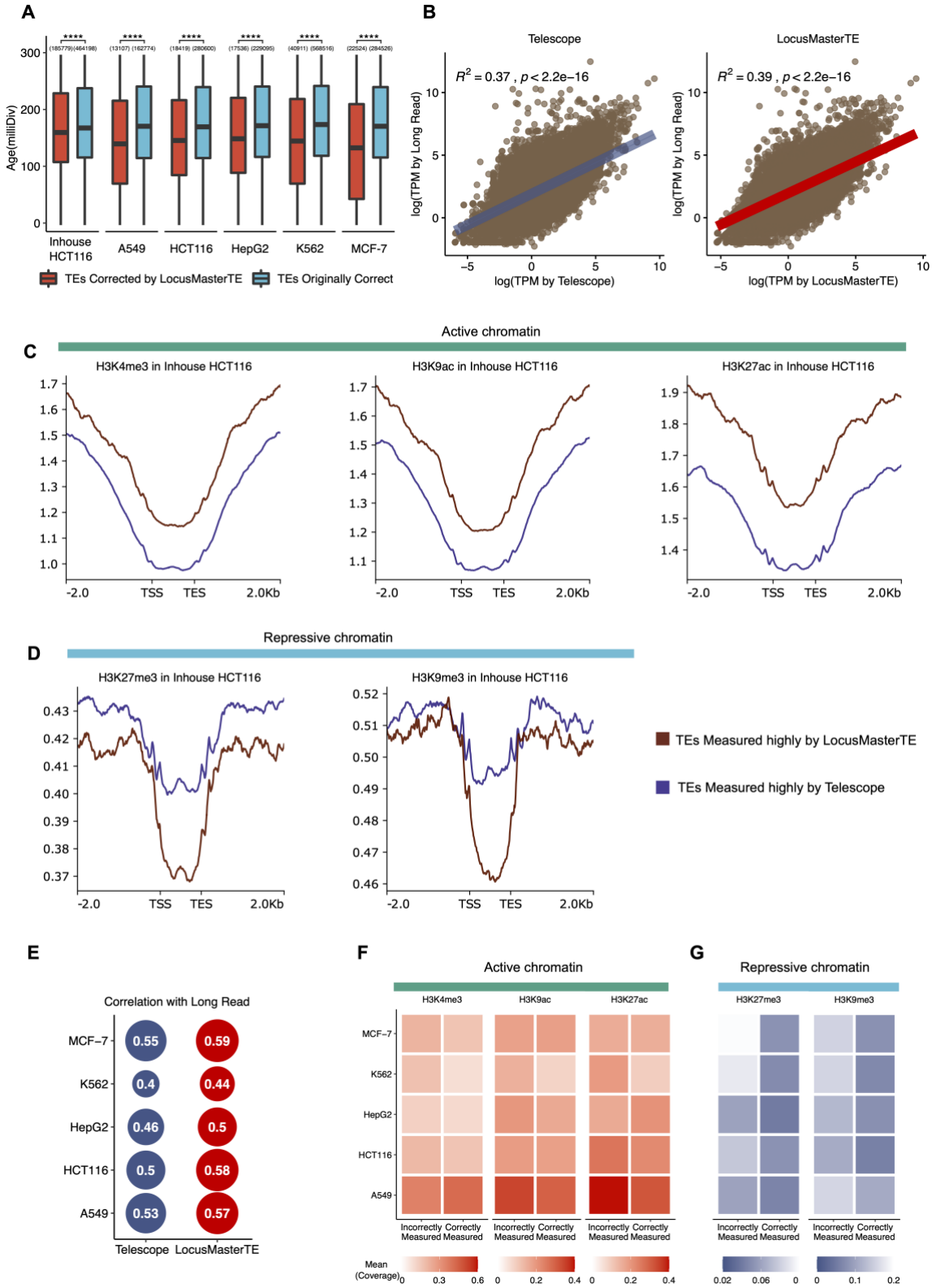

Figure 4: Identifying characteristics of quantified TEs using TE ages and histone

(A) Age distribution of TEs across all cell lines between two groups of TEs. Red: TEs quantified differently between LocusMasterTE and Telescope (Bendall et al. 2019) (TEs corrected by LocusMasterTE). Blue: TEs quantified the same across LocusMasterTE and Telescope (Bendall et al. 2019) (TEs originally correct). Quantitative numbers are labelled above each box. (B) Correlation plots for Telescope (Bendall et al. 2019) and LocusMasterTE with long-read in inhouse HCT116 cell line. log10 TPM counts are used. (C) Three active chromatins coverage of inhouse HCT116 cell for two TE groups. Brown: TEs quantified higher in LocusMasterTE (TEs measured highly by LocusMasterTE). Blue: TEs quantified higher in Telescope (Bendall et al. 2019) (TEs measured highly by Telescope). (D) Two repressive chromatins coverage of inhouse HCT116 cell sample for the same two TE groups. (E) Summary bubble plot for correlation with long-read in SG-NEx cell lines. The white number represents correlation values. (F) Summary heatmap for three active chromatins and two repressive chromatins in SG-NEx cell lines for the same two TE groups described in (C).

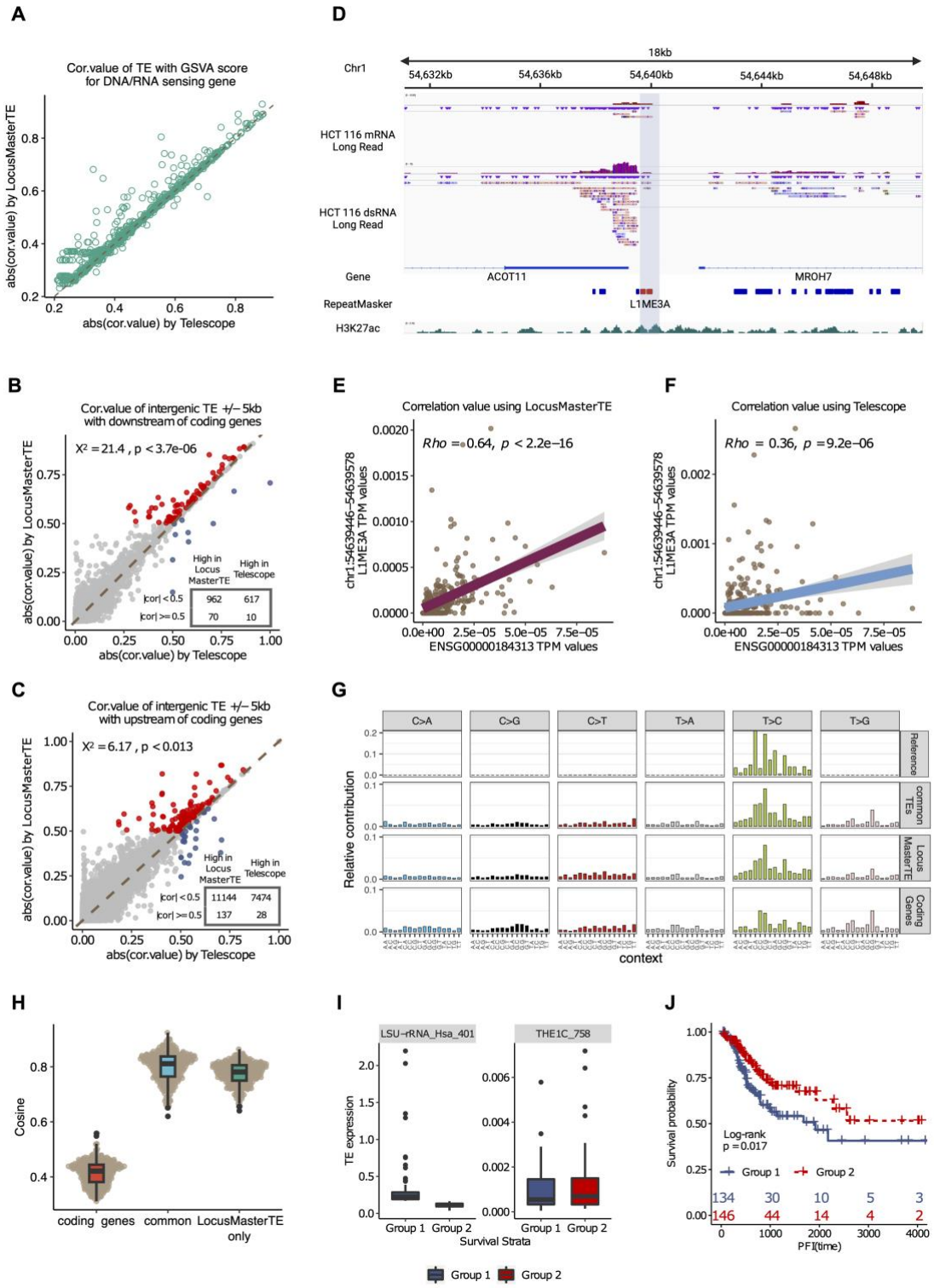

Figure 5: Connections of TE counts by LocusMasterTE in TCGA-COAD samples

(A) With two TE quantified matrices generated by Telescope (Bendall et al. 2019) and LocusMasterTE, comparison between absolute spearman correlation valueswith GSVA scores for DNA/RNA sensing genes. (B-C) Overview of correlation values between intergenic TEs and coding genes lying in intergenic TEs. We separate into two groups based on relative locations between TEs and correlated coding genes: downstream (B) and upstream (C). (D-F) Specific example showing statistically important relationship exclusively measured by LocusMasterTE. (D) IGV shows the presence of clear reads specifically for L1ME3a located in chromosome 1 from 54639446 to 54639578. (E) Correlation plot between MROH7 and one copy of L1ME3a located in chromosome 1 from 54639446 to 54639578 quantified by Telescope (Bendall et al. 2019). (F) Correlation plot between MROH7 and one copy of L1ME3a located in chromosome 1 from 54639446 to 54639578 quantified by LocusMasterTE. (G) Trinucleotide profiles for four groups: ADAR-WT-MDA5-5aza sample with inverted SINEs and ADAR1 (Mehdipour et al. 2020), TEs commonly detected for both Telescope (Bendall et al. 2019) and LocusMasterTE (common TEs), TEs solely detected by LocusMasterTE (LocusMasterTE only TEs), and coding genes (Coding genes) (H) Cosine similarity values between "common TEs (blue), "LocusMasterTE only TEs (green)", and "coding genes (red)" based on "Reference" for all TCGA-COAD tumor samples. (I) Expressions of two survival-related TEs in two clusters of TCGA-COAD patients. (J) Survival plot for two groups of patients separated using two selected TEs.