TE-Seq: A Transposable Element Annotation and RNA-Seq Pipeline

Maxfield M.G. Kelsey[1], Radha A. Kalekar[1], John M. Sedivy[1*]

[1]Center on the Biology of Aging, and the Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island, United States

*Corresponding author

Contact:

John M. Sedivy, Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island, United States. Email: john_sedivy@brown.edu

Maxfield M.G. Kelsey, Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island, United States. Email: maxfield_kelsey@brown.edu

Abstract

Motivation: The recognition that transposable elements (TEs) play important roles in many biological processes has elicited growing interest in analyzing sequencing data derived from this 'dark genome'. This is however complicated by the highly repetitive nature of these sequences in genomes, requiring the deployment of several problem-specific tools as well as the curation of appropriate genome annotations. This pipeline aims to make the analysis of TE sequences and their expression more generally accessible.

Results: The TE-Seq pipeline conducts an end-to-end analysis of RNA sequencing data, examining both genes and TEs. It implements the most current computational methods tailor-made for TEs, and produces a comprehensive analysis of TE expression at both the level of the individual element and at the TE clade level. Furthermore, if supplied with long-read DNA sequencing data, it is able to assess TE expression from non-reference (polymorphic) loci. As a demonstration, we analyzed proliferating, early senescent, and late senescent lung fibroblast RNA-Seq data, and created a custom reference genome and annotations for this cell strain using Nanopore sequencing data. We found that several retrotransposable element (RTE) clades were upregulated in senescence, which included non-reference, intact, and potentially active elements.

Availability and implementation: TE-Seq is made available as a Snakemake pipeline which can be obtained at https://github.com/maxfieldk/TE-Seq. All software dependencies besides Snakemake and Docker/Singularity are packaged into a container which is automatically built and deployed by the pipeline at runtime.

1 Introduction

Transposable elements (TEs) are mobile genetic elements that share a unique ability, using a variety of mechanisms, to move throughout genomes. These processes can increase TE genomic copy-number, so much so that TEs comprise roughly half of the human genome, and similarly large fractions of other mammalian and eukaryotic genomes (Almojil *et al.*, 2021). Their impact on biological processes is hard to overstate: they are involved in many fundamental cell-physiological processes, including chromosome-segregation, gene-regulation, and telomere function (Almojil *et al.*, 2021; Bourque *et al.*, 2018). Their less welcome participation in patho-physiological processes is also well documented, wherein they have been shown to promote chronic inflammation, tumorigenesis, and neurodegeneration (Frost and Dubnau, 2024; Gorbunova *et al.*, 2021; Mendez-Dorantes and Burns, 2023).

While the broad relevance of TEs is clear, the repetitive nature of their sequences imposes analytical challenges, and hence they are often neglected in RNA-Seq workflows and analyses (Lanciano and Cristofari, 2020). This pipeline aims to render TE investigation more tractable by addressing concerns pertaining to: i) imperfect alignment of repetitive elements, ii) non-reference elements, iii) non-autonomous transcription of TEs driven by adjacent genes, and iv) the quality of TE annotations.

Retrotransposable elements (RTEs), in particular Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs) and Endogenous Retroviruses (ERVs) are of considerable interest because their abundance and activity in many mammalian genomes. They are classified into superfamilies (e.g. LINE), families (e.g. LINE-1), and subfamilies (e.g. L1HS, specific for *Homo sapiens*).

RTE-count quantification has to confront the issue of multimapping reads: if two insertions in a genome are identical in sequence, it is impossible to uniquely assign reads derived from them to either of the loci. In practice, however, almost all RTE copies contain some polymorphic sites (such as SNPs or indels), allowing some fraction of their reads to be uniquely mappable. A strategy of examining only uniquely mapped reads may at first glance appear attractive, but this approach is biased against evolutionarily young RTE elements, which are less polymorphic than older elements (Kaul *et al.*, 2020; Teissandier *et al.*, 2019). However, the young RTE subfamilies are of greatest interest because they are the most mobile, and thus tools have been developed to deal with reads derived from non-uniquely mappable sequences.

This problem was recognized early on and programs such as TEtools randomly assigned a multi-mapper read to one locus; this greatly improved quantification accuracy (Lerat *et al.*, 2017). Other tools such as RepEnrich (Criscione *et al.*, 2014) focused on providing accurate assessments of expression at the TE subfamily level, by pooling all uniquely mapped reads to a given TE subfamily, and remapping multi-mappers to TE subfamily consensus sequences. As sequencing read lengths improved, with 150 bp paired-end reads becoming standard, the yield of uniquely mapping reads increased, and new strategies to deal with the remaining multi-mapping problem were developed, such as TElocal (https://github.com/mhammell-laboratory/TElocal), L1EM (McKerrow and Fenyö, 2020), and Telescope (Bendall *et al.*, 2019). These programs implemented iterative, expectation-maximization based assignments of multi-mappers to single loci. They rely on the assumption that loci which have uniquely mapped reads are likely truly expressed, and therefore multi-mapping reads are iteratively assigned to TE loci which accrue the largest number of uniquely mapped reads. Our TE-Seq pipeline utilizes the Telescope tool since it accepts user-provided, gene-transfer format (GTF) formatted TE annotations, making it

more flexible than either TElocal or L1EM; the former relies on pre-built custom-TE indices and the latter focuses solely on LINE-1 elements.

A complete account of every TE insertion in a genome would greatly increase the quality and interpretability of TE RNA-Seq results. Currently mobile RTE subfamilies, such as L1HS, are constantly evolving and generating new germline insertions. Each individual human has approximately 100 L1 insertions which are not captured by a given reference genome (e.g. Hg38). Analytically, this means that reads derived from such elements will necessarily be mistakenly assigned to other, reference elements. This problem is particularly acute since the polymorphic, non-reference RTEs are likely some of the most active elements in the genome. As a solution to this problem, our TE-Seq pipeline allows for the optional injection of Nanopore DNA sequencing data in order to call non-reference TE insertions and include them in downstream analyses.

Recognition of an element's 'genic context', or its spatial relationship to cellular genes, helps to estimate potential regulatory influences between TEs and the surrounding host genome, in particular, whether an element was autonomously expressed (transcribed by its own promoter) or whether it was passively transcribed by gene read-through is of interest. The latter possibility is diminished if no adjacent genes exist. To this end, the TE-Seq pipeline categorizes elements according to their relationship with cellular genes. This information serves another diagnostic function: if libraries differ greatly in terms of intronic content (a potential technical artifact owing to differences in sample preparation), this can lead to a spurious apparent increase in TE expression. The degree to which this pitfall may affect a dataset is assessed by comparing changes in intergenic-TE expression to changes in intronic-TE expression.

2 Implementation

2.1 Overall design

The TE-Seq pipeline consists of two overarching modules, *Annotate Reference (AREF)*, and *Short-read RNA-Seq (SRNA)*. Here we outline the overall workflow and in the next section we provide extended details for selected steps (see Fig. 1 for a schematic outline of TE-Seq).

The AREF module begins by fully annotating a user-provided genome for TE content and identifying functionally important groupings of elements. This enriched annotation set allows the subsequent analysis to probe distinctions in expression between truncated, full-length, and full-length open reading frame intact TEs, as well as to distinguish between intergenic versus intragenic elements. If provided with long-read Nanopore DNA-Seq data derived from the specimens under investigation, this pipeline will call non-reference TE insertions (polymorphic TEs not captured by the reference genome) and create an augmented reference genome containing sequences and associated annotations describing all non-reference TE element insertions. This enables the analysis of polymorphic TE elements, which are likely to include some of the most active elements. For use-case flexibility, the AREF module has a number of workflow modifying parameters. These parameters allow users to choose whether to create new annotations from scratch, to update annotations using long-read sequencing data, or to use existing annotations previously created during a previous run of the pipeline in another project.

Starting with raw sequencing reads, the SRNA module performs standard read-level quality control, genomic alignment, and quantification of gene counts. Repeat-specific tools are then deployed to quantify repeat element expression, either by using only uniquely mapped reads, or, by using the information provided by unique mappings to guide probabilistic assignment of multi-mapping reads to specific TEs. Differential transcript expression of

repetitive and non-repetitive elements is assessed using DESeq2. Gene set enrichment analyses are performed for both genes and families of repetitive elements.

2.2 Annotation set creation and refinement

Annotation curation begins with a user-provided reference genome. If a precomputed RepeatMasker.out file is not provided, RepeatMasker (www.repeatmasker.org) will be run on each chromosome to annotate repetitive elements. Next, this RepeatMasker.out file is converted to GTF format with a Bash script, and is then checked for errors and converted to GTF, GFF3 and BED file formats. User provided RefSeq gene annotations undergo a similar process, and ultimately RepeatMasker-derived and RefSeq-derived annotations are merged to create a TxDb R object.

Next, RepeatMasker-derived annotations are processed by R scripts to enable downstream analyses. By default, RepeatMasker splits annotations for fragmented repetitive elements. Adhering to 'tidy' data conventions, we consolidate these split annotations in an R data frame, taking note of their fragmented status. TEs are then further annotated on the basis of their inclusion in phylogenetic and functional classes. RTEs are grouped at the superfamily, family, and subfamily levels. Elements are denoted as full-length or truncated depending on whether they cover at least 95% of their representative consensus sequence. TE subfamilies are deemed 'young' if the average percent divergence (as determined by RepeatMasker) of all subfamily members is less than 3% for LINEs, 8% for SINEs, 12% for ERVs, and 15% for all others. These thresholds are user-tunable parameters, and are set by default, where possible, to denote as 'young' those subfamilies capable of retrotransposition (e.g. L1HS and L1PA2 for human

LINEs). Evolutionarily young, potentially active elements of two retrotransposon clades, LINEs and ERVs, are further examined below.

Young, potentially active L1 element sequences are analyzed for protein coding potential (in the human: L1HS and L1PA2; in the mouse: L1Md, L1Tf, L1 Gf and L1A subfamilies; in other species: the least diverged 7 LINE subfamilies). For each full-length sequence, open reading frames (ORFs) are identified. Not all ORFs in all elements will be intact, in fact, many will contain SNPs or indels as a result of genetic drift. To identify consensus ORFs, ORF length frequencies are tabulated, Z-scores are computed, and ORF lengths with a Z-score > 4 are selected for consensus sequence generation. All ORFs are subsequently aligned to these consensus sequences, and their percent amino acid difference is computed. ORFs which are less than 5% diverged are considered intact, and elements are considered intact if they possess a full complement of intact ORFs.

Endogenous retrovirus (ERV) elements are composed of long terminal repeat (LTR) sequences and internal (Int) sequences. RepeatMasker does not annotate whole proviruses, but instead splits them into LTRs and Ints. Most LTRs in genomes are present as ‘solo’ LTRs that do not flank an internal ERV sequence (solo LTRs result from host recombination between two LTRs of a provirus). The TE-Seq pipeline thus distinguishes between proviral and solo LTRs, as well as between 5’LTR-flanked and 5’LTR-deficient ERV sequences. For all LTRs, we determine whether they are flanked by full-length (FL) or truncated (Trnc) Ints (+/- 500bp), thereby generating the following categories: LTR (Solo), 5'LTR (Trnc Int), 3'LTR (Trnc Int), 5'LTR (FL Int), 3'LTR (FL Int). For all Ints, we determine whether these have a 5’LTR and mark them accordingly: Int (Has 5'LTR), or Int (No 5'LTR). Additionally, for young, potentially intact ORF encoding ERV element subfamilies (in the human: HERVK, HERVL, HERVW (Garcia-Montojo

et al., 2018; Jakobsson and Vincendeau, 2022); in the mouse: MMTV, ETn, IAP, MMVL, MMERGLN, MuRRs, MuLV, MERVL subfamilies (Stocking and Kozak, 2008); in other species: the least diverged 7 LTR subfamilies), we assign each Int and adjacent (+/- 500bp) LTRs of the same family a unique "proviral_group_id", such that their counts can be pooled and/or assessed for concordance.

All repetitive elements are examined for their proximity to non-repetitive genetic elements (such as cellular genes). The distance to the most proximal transcript is recorded, and the repetitive element is categorically denoted as either exonic, intronic, coding/non-coding gene proximal (+/- 10,000bp from transcript start/end), or intergenic.

2.3 Nanopore DNA-Seq

The TE-Seq pipeline allows optional non-reference TE insertion calling if provided with Nanopore DNA-Seq reads. Nanopore basecalling is performed with the Oxford Nanopore Dorado program, and reads are aligned to a reference genome using Minimap2 (Li, 2018). Long-read sequencing and alignment quality metrics are compiled using PycoQC (Leger and Leonardi, 2019) and Samtools stats (Danecek *et al.*, 2021). Aligned DNA reads are then fed to the TLDR program (Ewing *et al.*, 2020), which calls non-reference insertions by examining clusters of clipped-reads whose clipped regions align to a TE consensus sequence. We then filter these insertions to retain only those high-confidence calls which have passed all TLDR filters, and further require that: i) the median MapQ score of supporting reads be 60, ii) the insert has a target-site duplication (TSD) which is a hallmark of retrotransposition, and iii) the insert be supported by at least 10 reads of which 3 must fully span the insert. We then append the insert's consensus sequence (and 30bp of reference flanking region up and downstream of the insert) to

the provided reference genome where it becomes a standalone, mappable contig. RepeatMasker then annotates these appended contigs for repetitive element content, and these non-reference insertion annotations are merged with the reference TE RepeatMasker set for further processing as described in the previous section. If provided with a species-specific database of TE polymorphisms, TLDR annotates whether insertions are 'known' or 'novel'. TLDR provides such databases for human and mouse (see (Ewing *et al.*, 2020), Table S4 for more information on the 17 datasets comprising this database). Additionally, we optionally call SNPs using the PEPPER pipeline (Shafin *et al.*, 2021), and incorporate high-quality SNP calls into the reference genome allowing for more accurate mapping of reads overlapping these polymorphisms.

If Nanopore DNA-sequences are not available, the problem of non-reference TEs can nevertheless be attenuated by using the most up-to-date reference genomes, such as telomere-to-telomere assemblies (Nurk *et al.*, 2022) which provide a more complete map of TE insertions, especially in hard to assemble regions such as centromeres.

2. 4 RNA-Seq analysis

Sequencing reads are trimmed using fastp (Chen *et al.*, 2018) and aligned to the (*custom*) reference genome reference with STAR (Dobin *et al.*, 2013), allowing for up to 100 multiple mapping locations per read. This allowance for multimappers aids in accurately assigning TE-aligned reads to the correct TE locus. Gene counts are produced using featureCounts (Liao *et al.*, 2014) and the user-provided RefSeq transcript annotation (in the present case, the October 2023 revision, GCF_009914755.1). Repetitive element count estimates are produced with the Telescope tool (Bendall *et al.*, 2019) and the custom RepeatMasker derived annotation resulting from the AREF module. This tool allows for multi-mapping reads overlapping RTEs to be

assigned to the best supported locus via an iterative expectation maximization (EM) algorithm (see Telescope methods). We obtain two sets of TE counts: one conservative estimate of locus expression which allows only for uniquely-mapped reads to contribute to total counts, and one more relaxed (and potentially accurate, as claimed in (Bendall *et al.*, 2019)) set of counts which incorporates multi-mapping reads. This inclusion of multi-mapping reads is particularly important for young TE elements, which are highly similar in sequence and contain long stretches of non-uniquely mappable sequences. While overall a less biased strategy than only considering uniquely mapping reads (Bendall *et al.*, 2019), EM re-assignment of multi-mapping reads is imperfect, and element-level read estimates should be interpreted as best guesses when they are comprised of a substantial fraction of multi-mapping reads.

Count normalization for genes and repetitive elements is performed using DESeq2's 'median of ratios' method (Love *et al.*, 2014). Size factors are estimated by setting the 'controlGenes' parameter of the 'estimateSizeFactors' factors to include only RefSeq genes. This is done so as to prevent the very large number of lowly-expressed repetitive elements in a RepeatMasker-derived annotation from adversely biasing median-based count normalization. Differential expression is then assessed using DESeq2's negative binomial linear modeling. Batch effect correction is implemented by introducing a batch variable in the sample specification file. If sequencing data contain batch effects, one must be mindful that batch effect correction is an imperfect procedure which can in itself introduce a bias (Leek *et al.*, 2010; Nygaard *et al.*, 2016). This problem is particularly acute when biological conditions of interest are imbalanced with respect to batch. The TE-Seq pipeline accounts for batch effects in two places: the first is in DESeq2's linear model, and the second is during the production of batch-corrected count tables using limma (Ritchie *et al.*, 2015). DESeq2's linear modeling will better

handle the batch effect, and its p-value estimates (for individual genes and repetitive elements) will be more reliable than any statistics produced downstream of limma 'batch-corrected' counts (comparisons between repetitive element family total counts). Therefore, in the context of an unbalanced dataset, TE family comparisons based on adjusted counts need to be interpreted with care, and in the light of non-batch adjusted results. For a more thorough discussion of the issue, readers may consider Nygaard *et al.* (2016).

Normalized counts and differential expression results are next analyzed with a set of R scripts. Example products of this analysis are shown in the results section. Analyses at both the individual gene/TE level, as well as at the gene set/TE family level are performed. TE groupings are assessed for enrichment using both gene set enrichment analysis (GSEA) and pooled-count strategies. GSEA can detect small, concerted changes in expression across a family of related TEs. Normalized TE count pooling at the family or functional level (and subsequent assessment for differential expression by means of a t-test) allows for a TE family to be counted as significantly overexpressed if only a small number of its members are dramatically overexpressed.

Correlations in expression between adjacent gene-TE pairs are established to shed light on potential bidirectional gene-regulatory influences. Read-mapping to select TE families is also examined by means of composite bigwig tracks, which show read density across a consensus element's length, revealing potential biases in alignment to particular TE sequence regions/features.

Finally, the pipeline produces an html report consolidating the most insightful visualizations. Plots are stored as PDFs such that they can be resized and modified as needed in vector graphics software such as Adobe Illustrator.

3 Results

In order to catalogue all non-reference RTE insertions in LF1 human primary lung fibroblast cells (Brown *et al.*, 1997), we performed Nanopore DNA sequencing. Sequencing coverage and N50 values were high (>60X, >24 kb). We called non-reference insertions using TLDR, and retained only those insertions passing both TLDR and our in-house quality filters (2.3 above). We observed 226 Alu insertions, 117 L1 insertions, and 24 SVA insertions (Fig. 2A). While some L1 insertions were truncated, we found 65 full-length L1HS insertions, of which 46 were intact with conserved ORFs. Most insertions were intergenic or intronic (Fig. 2B). A smaller fraction were either adjacent to coding or non-coding genes, or found within an exon or a non-coding transcript . Compared to reference L1HS elements, non-reference L1HS elements were much more likely to be intact, and non-reference elements made up 30% (46/151) of all intact L1HS elements (Fig 2C). When non-reference intact L1HS consensus sequences were phylogenetically compared to reference intact L1HS elements, non-reference insertions were dispersed throughout the phylogeny, but with an increased prevalence among the younger elements (Fig. 2D). Assuming that the most phylogenetically proximal reference element is the likeliest non-reference progenitor, a small number of reference intact L1HS elements were predicted to account for the majority of non-reference insertions (Fig. 2E). Under this assumption, the L1HS element (17_q24.2_2) produced eight novel insertions, indicating it may be a particularly active element in the germline. We also generated a pie chart showing the genome fraction occupied by repetitive elements in this genome (Fig. 2F).

Starting with the near-complete T2T-HS1 human reference genome (Nurk *et al.*, 2022), we then created a custom LF1 genome which incorporates quality-filtered TE insertions,

rendering them alignable and hence open to scrutiny by RNA-Seq analysis. Enhanced TE annotations were then created to facilitate the thorough investigation of these elements. We applied our pipeline to RNA-Seq data from our laboratory on LF1 cells spanning three biological conditions. LF1 cells were passaged to replicative senescence, and were harvested 4 and 16 weeks later, yielding early senescence (ESEN) and late senescence (LSEN) conditions. Early passage proliferating LF1 cells (PRO) were used as a control. Poly-A selected, stranded, 150bp paired-end libraries were sequenced in quadruplicate.

Principal component analysis revealed samples mostly clustered by biological condition (Fig. 3A), excepting one ESEN sample, which clustered closer to LSEN samples. Senescence status neatly divided the first principal component, which accounts for 72.63% of total variation. We observed 7341/6245 up/down-regulated genes in ESEN and 7274/6117 up/down-regulated genes in LSEN (Fig. 3B). While most differentially expressed genes were shared between both senescent conditions, roughly 40% of DEGs were unique to one condition, suggesting biological differences between early and late senescence. Hallmark gene set enrichments by GSEA (Fig. 3C) were consistent with previous reports demonstrating heightened inflammatory signatures and P53 pathway activity, as well as diminished activity of pathways associated with proliferation, including E2F targets, and G2M checkpoint (Hernandez-Segura *et al.*, 2018). Canonical transcriptional markers of senescence were significantly elevated in ESEN and LSEN, including CCL2, CDKN1A (P21), CDKN2A (P16), and IL6 (Fig. 3D). The SenMayo gene set was enriched in both ESEN and LSEN (Fig. 3E). Interestingly, hierarchical clustering of SenMayo gene expression showed that this signature was dynamic in senescence, with some SenMayo genes being differentially expressed between ESEN and LSEN (Fig. 3F). The direction of most changes observed in ESEN were preserved in LSEN, but their magnitude was modulated in

LSEN. One large block of genes (top of panel) was highly upregulated in ESEN and comparably modestly upregulated in LSEN. While STRING analysis did not detect functional enrichments in this block compared to the broader SenMayo set, many of these genes were cytokines (IL1A, IL1B, and IL6) and chemokines (CCL2, CXCL8, CXCL10). Another large block (middle of panel) was upregulated modestly in ESEN and much more strongly in LSEN.

The majority of differentially expressed repetitive elements were upregulated: 10204 were upregulated in ESEN and 7418 were upregulated in LSEN, while 4766 were downregulated in ESEN and 6206 were downregulated in LSEN. Nearly half of the up- and downregulated TEs were shared between senescent conditions (Fig. 4A). At the superfamily level, RTEs (LINEs, SINEs, LTRs) and DNA transposons were upregulated, while retroposons and satellites were downregulated (Fig. 4B). An examination at the RTE family level found that L1 elements were particularly highly expressed during early and late senescence (Fig. 4C). This observation was further borne out upon examining expression at the RTE subfamily level, where evolutionarily recent L1 subfamilies saw the largest expression fold changes in early and late senescence (Fig. 4D). Gene set enrichment analysis of these RTE subfamilies (Fig 4F), in which subfamilies were treated as gene sets, was concordant with the log2 fold change aggregated count analysis (Fig. 4D). Both showed that evolutionarily recent L1 and HERV-int subfamilies were enriched in ESEN and LSEN, while SVA sets were depleted (Fig. 4D, F).

Focusing on L1, we observed that many loci spanning all L1 clades were differentially expressed, with the majority being upregulated (Fig. 4E). Most counts were derived from ancient, truncated elements, particularly in the L1M group. This is in keeping with the fact that the majority of L1 elements belong to this group, and reflects the overall loss of heterochromatin in senescence. Of particular interest, however, are the full-length and intact L1HS elements,

which retain biological activities, including cDNA generation and retrotransposition (Fig. 4G). These elements had significantly elevated transcript counts in both ESEN and LSEN (Fig. 4G). Changes in full-length L1HS element expression were not correlated with expression changes in adjacent cellular genes (Fig. 4J), suggesting that they are transcribed from their own promoters. Analysis of full-length L1HS elements at the locus level revealed two significantly upregulated elements, both of which were intergenic, and one of which (14q23.2_3) was intact (Fig. 4I).

Read alignments to full-length L1HS elements spanned the entire element and did not exhibit a 3' bias, which would indicate the misattribution of truncated element reads to full-length elements (Fig. 4H4H). Seeing as many TEs reside in introns, differences in intronic read fraction, be this due to differences in intron retention or sequencing library preparation, could lead to spurious conclusions about TE regulation (Faulkner, 2023). We did not observe significant changes in the amount of intronic signal between conditions (Fig. 4K), in fact, intronic signal was decreased in senescent cells. There was, however, a small and significant increase in intergenic signal in both senescent conditions (Fig. 4K). Finally, ESEN showed a small increase in the overall share of the transcriptome derived from repetitive elements (Fig 4L). Taken together, these data document an upregulation of TEs, including young, potentially active RTEs, in ESEN and LSEN conditions.

4. Discussion

Here we present a pipeline for the analysis of TE RNA-Seq data. Our aim was to render an investigation into these elements more tractable to the non-expert. We address known pitfalls in the transcriptomic analysis of these elements by generating an enhanced set of TE annotations, employing up-to-date TE-minded computational methods, and discriminating between signals

which originate from autonomous transcripts versus elements transcribed from neighboring genes. To this end, our pipeline makes the following distinctions between elements: genic versus non-genic, evolutionarily young versus old, truncated versus full-length, and intact versus open reading frame (ORF) disrupted. The pipeline can be employed with uniquely mapping reads only, or it can utilize expectation maximization assignments, and given sufficient coverage individual TE loci can be queried for expression. Furthermore, if provided with Nanopore DNA sequences, we provide the option to build a custom reference genome which includes all polymorphic, evolutionarily young elements not included in reference genomes. Taken together, these measures can provide a more accurate assessment of TE expression than is currently possible with standard tools, and should improve our ability to assess whether TEs are contributing meaningfully to a biological problem of interest. As a demonstration, we applied this pipeline to a new dataset from primary human fibroblasts and showed that full-length and intact young L1 elements were significantly upregulated in replicative senescence. This is consistent with previous data from our and other groups. Of particular interest was the finding that numerous non-reference L1 elements, which were previously not visible to us, contribute to this upregulation.

4.1 Experimental considerations

Several experimental considerations can help to produce a high quality input dataset for this pipeline. 1) Sequencing read mappability is a function of read-length: a 50bp single-end library will have a much greater number of multi-mapping reads as compared to a 150bp paired-end library (Teissandier *et al.*, 2019). Insofar as TE expression analysis is an important goal, we recommend opting for the longest reads possible. 2) Young L1 subfamilies contain a G-rich

stretch of nucleotides in their 3' UTR which is thought to form a G-quadruplex secondary structure (Sahakyan *et al.*, 2017). Polymerases used during sequencing library preparation can struggle with this leading to artificially depressed counts. Utilizing reverse transcriptase enzymes with high-processivity tolerant of RNA secondary structures, such as those encoded by group II introns, is therefore preferable. Alternatively, PCR-free approaches such as long-read Nanopore direct RNA sequencing may provide a more accurate assessment of young L1 transcript abundance.

4.2 Software requirements

This pipeline uses the Snakemake (Mölder *et al.*, 2021) workflow manager, and consists of several parts. A main snakefile orchestrates the workflow logic and deploys module level snakefiles which in turn contain the rules that specify each step of the analysis. Conda must be available so as to create a Snakemake containing environment. All software dependencies besides Snakemake and Docker/Singularity are packaged into a container which is automatically built and deployed by the pipeline at runtime. Users can alternatively choose to manually build the several conda environments required by the pipeline using the provided yaml environment specifications (this will however be slower and less reproducible). This pipeline was developed on a compute cluster running a RedHat Linux OS and which uses the SLURM workload manager. Nevertheless the use of Docker containers should enable users on other operating systems to run the pipeline without issue.

4.3 Hardware requirements

This pipeline allows for parallel execution of jobs which can occur simultaneously. Consequently, it is highly recommended to execute this pipeline on a compute cluster to take advantage of the parallelization offered by Snakemake. Snakemake is designed to work with many commonly-used cluster workload managers such as SLURM. We provide a default Snakemake profile which is compatible with SLURM, and can be straightforwardly modified for use with other workload managers. Many steps require a substantial amount of RAM (>20 GB) to be available on the system. The Docker container is built to work with both X86 and ARM based CPU architectures.

References

Almojil, D. *et al.* (2021) The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes. *Genes*, **12**, 918.

Bendall, M.L. *et al.* (2019) Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLOS Computational Biology*, **15**, e1006453.

Bourque, G. *et al.* (2018) Ten things you should know about transposable elements. *Genome Biology*, **19**, 199.

Brown, J.P. *et al.* (1997) Bypass of senescence after disruption of p21CIP1/WAF1 gene in normal diploid human fibroblasts. *Science*, **277**, 831-834.

Chen, S. *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884-i890.

Criscione, S.W. *et al.* (2014) Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, **15**, 583.

Danecek, P. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.

Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.

Ewing, A.D. *et al.* (2020) Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Molecular Cell*, **80**, 915-928.e915.

Frost, B. and Dubnau, J. (2024) The Role of Retrotransposons and Endogenous Retroviruses in Age-Dependent Neurodegenerative Disorders. *Annu Rev Neurosci*, **47**, 123-143.

Garcia-Montojo, M. *et al.* (2018) Human endogenous retrovirus-K (HML-2): a comprehensive review. *Critical Reviews in Microbiology*, **44**, 715-738.

Gorbunova, V. *et al.* (2021) The role of retrotransposable elements in ageing and age-associated diseases. *Nature*, **596**, 43-53.

Hernandez-Segura, A. *et al.* (2018) Hallmarks of Cellular Senescence. *Trends in Cell Biology*, **28**, 436-453.

Jakobsson, J. and Vincendeau, M. (2022) SnapShot: Human endogenous retroviruses. *Cell*, **185**, 400-400.e401.

Kaul, T. *et al.* (2020) Comparative analysis on the expression of L1 loci using various RNA-Seq preparations. *Mob DNA*, **11**, 2.

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

Leek, J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, **11**, 733-739.

Leger, A. and Leonardi, T. (2019) pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, **4**, 1236.

Lerat, E. *et al.* (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.

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

Liao, Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923-930.

Love, M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

McKerrow, W. and Fenyö, D. (2020) L1EM: a tool for accurate locus specific LINE-1 RNA quantification. *Bioinformatics*, **36**, 1167-1173.

Mendez-Dorantes, C. and Burns, K.H. (2023) LINE-1 retrotransposition and its deregulation in cancers: implications for therapeutic opportunities. *Genes Dev*, **37**, 948-967.

Mölder, F. *et al.* Sustainable data analysis with Snakemake. In.: F1000Research; 2021.

Nurk, S. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44-53.

Nygaard, V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, **17**, 29-39.

Ritchie, M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

Sahakyan, A.B. *et al.* (2017) G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol*, **24**, 243-247.

Shafin, K. *et al.* (2021) Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*, **18**, 1322-1332.

Stocking, C. and Kozak, C.A. (2008) Endogenous retroviruses. *Cell Mol Life Sci*, **65**, 3383-3398.

Teissandier, A. *et al.* (2019) Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mobile DNA*, **10**, 52.

Acknowledgements

The authors would like to thank Jess Anderson for lab management and technical assistance, and members of the Sedivy lab and Center on the Biology of Aging for feedback and support. We are grateful to the Brown Center for Computation and Visualization (CCV) team for their management of the OSCAR high performance computing cluster which was used throughout this work.

Author Contributions

M.M.K. developed all code and figures, and performed long-read Nanopore DNA sequencing of LF1 cells. R.A.K. provided short-read RNA sequencing of LF1 cells. J.M.S. supervised the project. M.M.K. and J.M.S authored the manuscript.

Conflict of interest

J.M.S. is a cofounder and SAB chair of Transposon Therapeutics, and holds equity in PrimeFour and Atropos Therapeutics.

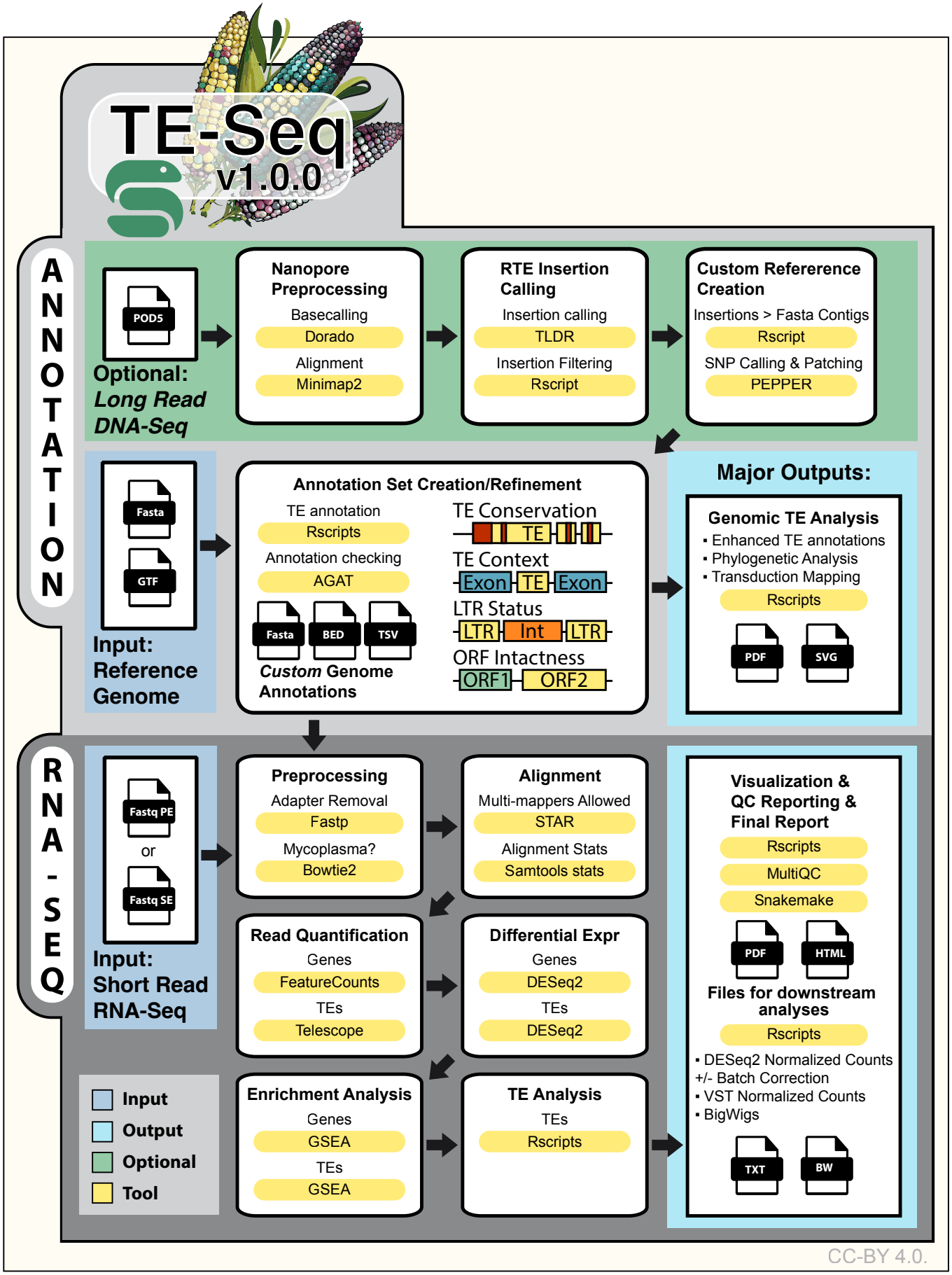
Funding

This work was supported by National Institutes of Health (NIH) grants R01 AG016694 and P01 AG051449 to JMS.

Code and data availability

Instructions for pipeline installation, configuration, and deployment are documented at the project GitHub page: https://github.com/maxfieldk/TE-Seq. This page also provides instructions

for the acquisition of human and mouse reference genome annotations. Nanopore DNA-Seq data have been deposited in the Gene Expression Omnibus (GEO) with accession number GSE268488. RNA-seq data are available upon request. Down-sampled versions of these data are made available for pipeline testing purposes and can be found at the GitHub page. Any other data or information relevant to this study are available from the corresponding author upon reasonable request.


TE-Seq
v1.0.0
ANNOTATION
POD5
Optional:
Long Read
DNA-Seq
Nanopore Preprocessing
Basecalling
Dorado
Alignment
Minimap2
RTE Insertion Calling
Insertion calling
TLDR
Insertion Filtering
Rscript
Custom Refererence Creation
Insertions > Fasta Contigs
Rscript
SNP Calling & Patching
PEPPER
Fasta
GTF
Input:
Reference
Genome
Annotation Set Creation/Refinement
TE annotation
Rscripts
Annotation checking
AGAT
Fasta
BED
TSV
Custom Genome
Annotations
TE Conservation
TE
TE Context
Exon
TE
Exon
LTR Status
LTR
Int
LTR
ORF Intactness
ORF1
ORF2
Major Outputs:
Genomic TE Analysis
▪ Enhanced TE annotations
▪ Phylogenetic Analysis
▪ Transduction Mapping
Rscripts
PDF
SVG
RNA-SEQ
Fastq PE
or
Fastq SE
Input:
Short Read
RNA-Seq
Preprocessing
Adapter Removal
Fastp
Mycoplasma?
Bowtie2
Alignment
Multi-mappers Allowed
STAR
Alignment Stats
Samtools stats
Read Quantification
Genes
FeatureCounts
TEs
Telescope
Differential Expr
Genes
DESeq2
TEs
DESeq2
Enrichment Analysis
Genes
GSEA
TEs
GSEA
TE Analysis
TEs
Rscripts
Visualization &
QC Reporting &
Final Report
Rscripts
MultiQC
Snakemake
PDF
HTML
Files for downstream
analyses
Rscripts
▪ DESeq2 Normalized Counts
+/- Batch Correction
▪ VST Normalized Counts
▪ BigWigs
TXT
BW
Input
Output
Optional
Tool

CC-BY 4.0.

Fig. 1. TE-Seq pipeline overview.

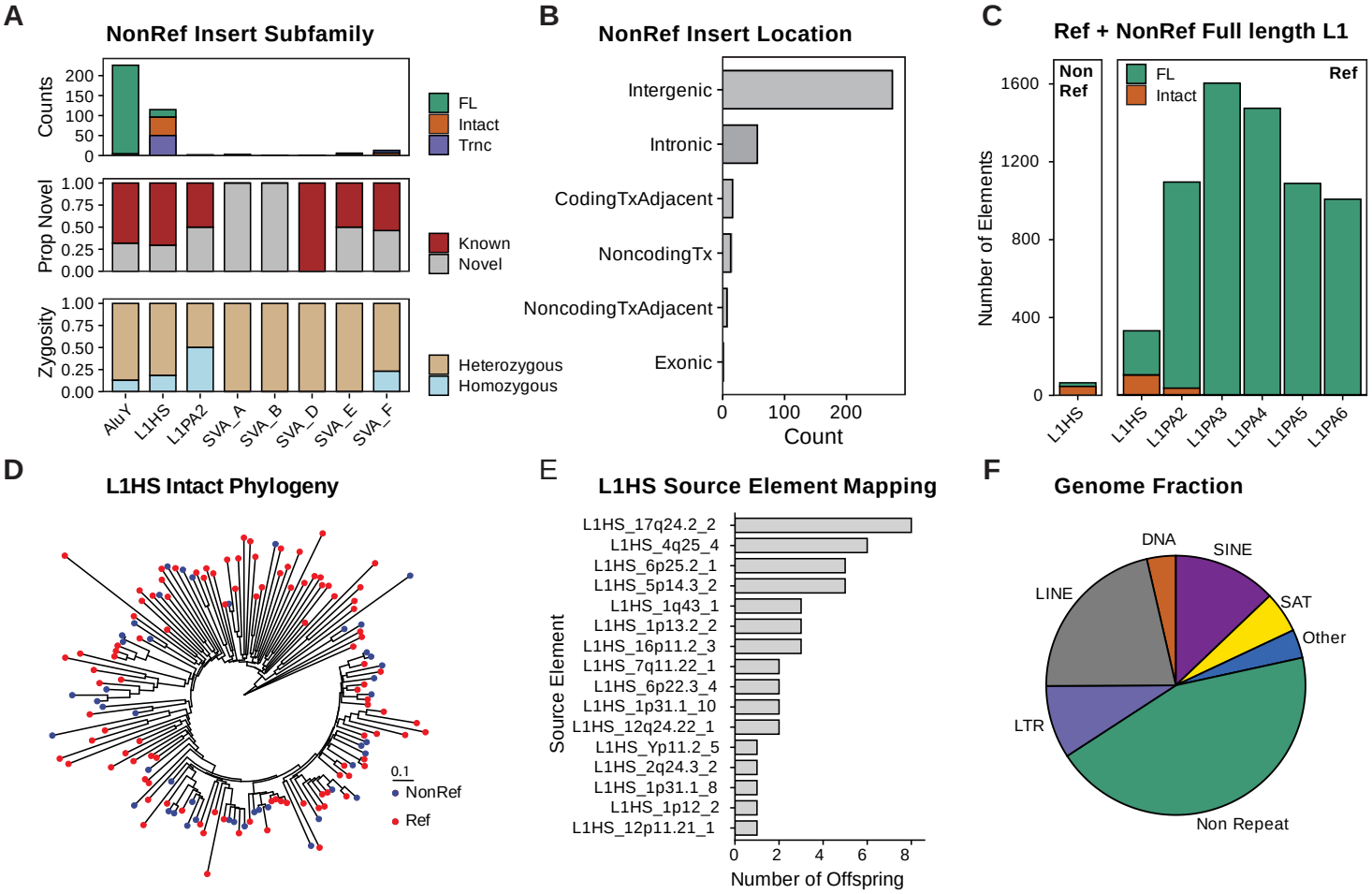

A
NonRef Insert Subfamily
Counts
200
150
100
50
0
FL
Intact
Trnc
Prop Novel
1.00
0.75
0.50
0.25
0.00
Known
Novel
Zygosity
Heterozygous
Homozygous
AluY
L1HS
L1PA2
SVA_A
SVA_B
SVA_D
SVA_E
SVA_F
B
NonRef Insert Location
Intergenic
Intronic
CodingTxAdjacent
NoncodingTx
NoncodingTxAdjacent
Exonic
0
100
200
Count
C
Ref + NonRef Full length L1
Non Ref
Ref
FL
Intact
Number of Elements
1600
1200
800
400
0
L1HS
L1HS
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
D
L1HS Intact Phylogeny
0.1
NonRef
Ref
E
L1HS Source Element Mapping
Source Element
L1HS_17q24.2_2
L1HS_4q25_4
L1HS_6p25.2_1
L1HS_5p14.3_2
L1HS_1q43_1
L1HS_1p13.2_2
L1HS_16p11.2_3
L1HS_7q11.22_1
L1HS_6p22.3_4
L1HS_1p31.1_10
L1HS_12q24.22_1
L1HS_Yp11.2_5
L1HS_2q24.3_2
L1HS_1p31.1_8
L1HS_1p12_2
L1HS_12p11.21_1
0
2
4
6
8
Number of Offspring
F
Genome Fraction
DNA
SINE
LINE
SAT
Other
LTR
Non Repeat

Fig. 2. | Nanopore DNA-Seq characterization of non-reference RTE insertions in a human genome. TE insertions in LF1 cells were called relative to the T2T-HS1 human reference genome using TLDR. (**A**) Insertion counts for Alu, L1 and SVA subfamilies (no NonRef ERVs were found). FL, full length; Intact, intact ORFs (for L1 only); Trnc, truncated. Insertions denoted as 'known' have been previously identified in human populations. (**B**) Positional relationships to cellular genes (for all NonRef RTEs in panel (A). 5' and 3' UTRs are included in the exonic category. Tx, transcript. (**C**) Reference versus non-reference full length L1HS and L1PA2 elements. (**D**) Phylogeny of intact L1HS elements. Tips are colored according to reference status. (**E**) Source elements were inferred for all non-reference L1HS insertions on the basis of phylogenetic relationships, and the number of offspring were tabulated for each source element. (**F**) Genomic occupancy for major clades of repetitive sequences. DNA, DNA transposons; SAT, satellite sequences.

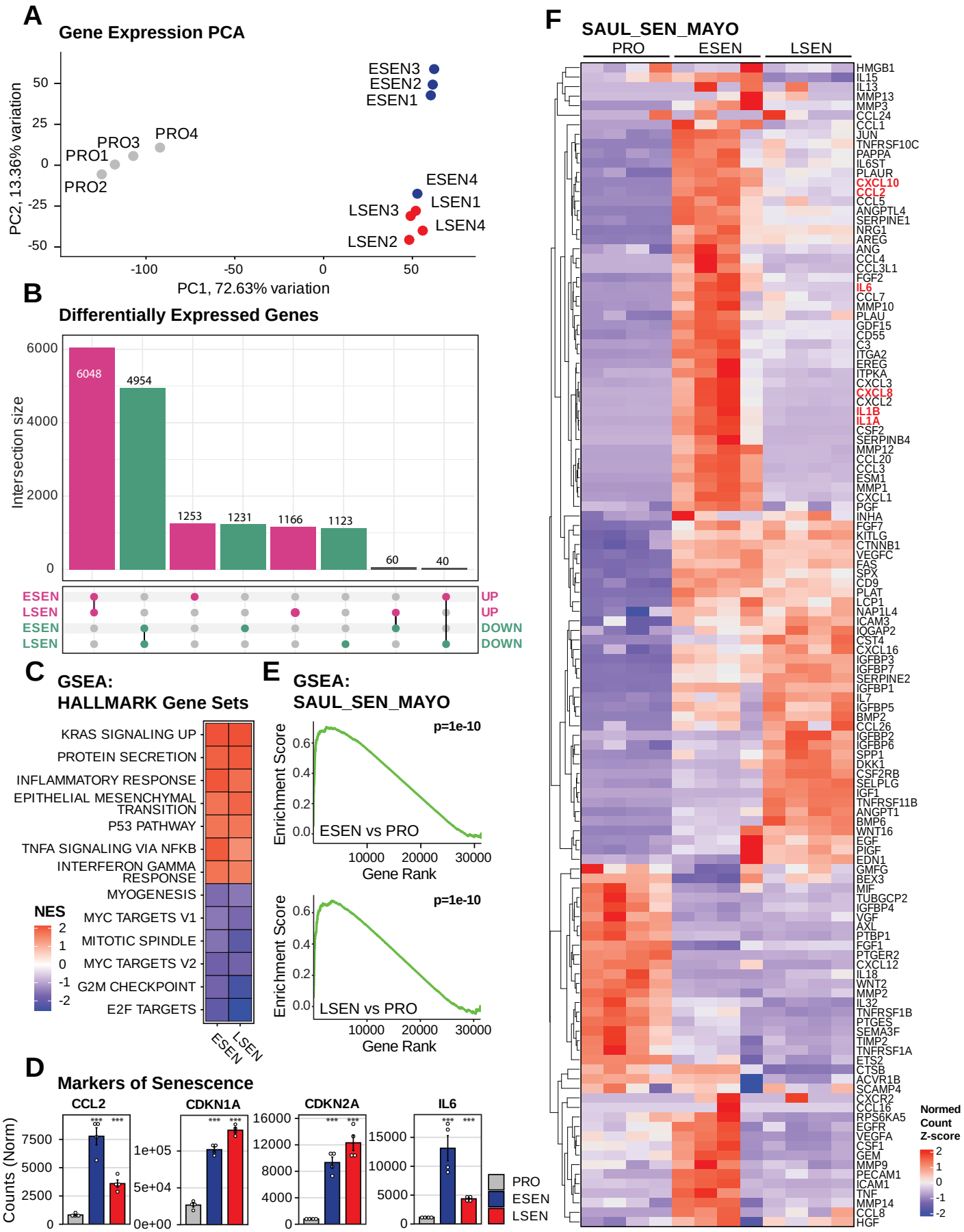

A
Gene Expression PCA
PC2, 13.36% variation
PC1, 72.63% variation
ESEN3
ESEN2
ESEN1
PRO4
PRO3
PRO1
PRO2
ESEN4
LSEN1
LSEN3
LSEN4
LSEN2
B
Differentially Expressed Genes
Intersection size
6048
4954
1253
1231
1166
1123
60
40
ESEN
LSEN
ESEN
LSEN
UP
UP
DOWN
DOWN
C
GSEA:
HALLMARK Gene Sets
KRAS SIGNALING UP
PROTEIN SECRETION
INFLAMMATORY RESPONSE
EPITHELIAL MESENCHYMAL TRANSITION
P53 PATHWAY
TNFA SIGNALING VIA NFKB
INTERFERON GAMMA RESPONSE
MYOGENESIS
MYC TARGETS V1
MITOTIC SPINDLE
MYC TARGETS V2
G2M CHECKPOINT
E2F TARGETS
NES
ESEN
LSEN
E
GSEA:
SAUL_SEN_MAYO
p=1e-10
Enrichment Score
ESEN vs PRO
Gene Rank
p=1e-10
LSEN vs PRO
D
Markers of Senescence
CCL2
CDKN1A
CDKN2A
IL6
Counts (Norm)
PRO
ESEN
LSEN
F
SAUL_SEN_MAYO
PRO
ESEN
LSEN
HMGB1
IL15
IL13
MMP13
MMP3
CCL24
CCL1
JUN
TNFRSF10C
PAPPA
IL6ST
PLAUR
CXCL10
CCL2
CCL5
ANGPTL4
SERPINE1
NRG1
AREG
ANG
CCL4
CCL3L1
FGF2
IL6
CCL7
MMP10
PLAU
GDF15
CD55
C3
ITGA2
EREG
ITPKA
CXCL3
CXCL8
CXCL2
IL1B
IL1A
CSF2
SERPINB4
MMP12
CCL20
CCL3
ESM1
MMP1
CXCL1
PGF
INHA
FGF7
KITLG
CTNNB1
VEGFC
FAS
SPX
CD9
PLAT
LCP1
NAP1L4
ICAM3
IQGAP2
CST4
CXCL16
IGFBP3
IGFBP7
SERPINE2
IGFBP1
IL7
IGFBP5
BMP2
CCL26
IGFBP2
IGFBP6
SPP1
DKK1
CSF2RB
SELPLG
IGF1
TNFRSF11B
ANGPT1
BMP6
WNT16
EGF
PIGF
EDN1
GMFG
BEX3
MIF
TUBGCP2
IGFBP4
VGF
AXL
PTBP1
FGF1
PTGER2
CXCL12
IL18
WNT2
MMP2
IL32
TNFRSF1B
PTGES
SEMA3F
TIMP2
TNFRSF1A
ETS2
CTSB
ACVR1B
SCAMP4
CXCR2
CCL16
RPS6KA5
EGFR
VEGFA
CSF1
GEM
MMP9
PECAM1
ICAM1
TNF
MMP14
CCL8
HGF
Normed Count Z-score

Figure 3 | RNA-Seq analysis of cellular genes. Human LF1 cells were passaged until replicative senescence and harvested 4 (ESEN) or 16 (LSEN) weeks later. Early passage proliferating (PRO) LF1 cells were used as the control. (**A**) PCA bi-plot of cellular genes. (**B**) Upset plot of differentially expressed genes. (**C)** Gene set enrichment analysis (GSEA) of MsigDb Hallmark gene set collection for ESEN vs PRO and LSEN vs PRO. (**D**) Bar plots showing differentially expressed markers of senescence. (**E**) GSEA plots of the SenMayo gene set for ESEN vs PRO and LSEN vs PRO. (**F**) Heatmap showing expression Z-scores for all SenMayo gene set members. Genes referred to in the main text are bolded in red. Unless otherwise stated, statistical significance was assessed by means of a two-sided t-test, with panel-level FDR corrected p-values >= 0.05 *, >=0.01 **, and >=0.001 ***.

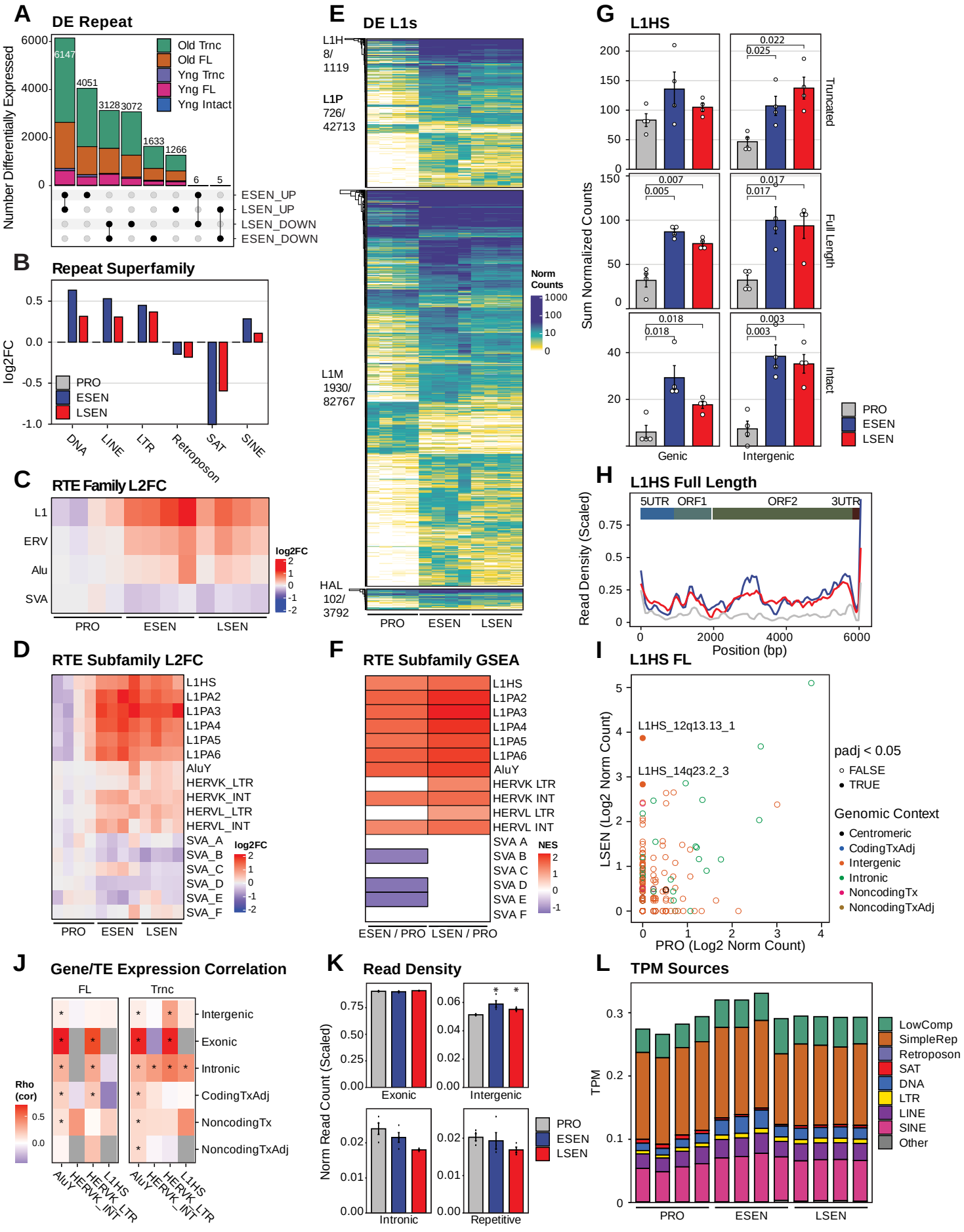

A
DE Repeat
Number Differentially Expressed
Old Trnc
Old FL
Yng Trnc
Yng FL
Yng Intact
6147
4051
3128
3072
1633
1266
6
5
ESEN_UP
LSEN_UP
LSEN_DOWN
ESEN_DOWN
B
Repeat Superfamily
log2FC
PRO
ESEN
LSEN
DNA
LINE
LTR
Retroposon
SAT
SINE
C
RTE Family L2FC
L1
ERV
Alu
SVA
log2FC
PRO
ESEN
LSEN
D
RTE Subfamily L2FC
L1HS
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
AluY
HERVK_LTR
HERVK_INT
HERVL_LTR
HERVL_INT
SVA_A
SVA_B
SVA_C
SVA_D
SVA_E
SVA_F
log2FC
PRO
ESEN
LSEN
E
DE L1s
L1H
8/
1119
L1P
726/
42713
L1M
1930/
82767
HAL
102/
3792
Norm
Counts
PRO
ESEN
LSEN
F
RTE Subfamily GSEA
L1HS
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
AluY
HERVK LTR
HERVK INT
HERVL LTR
HERVL INT
SVA A
SVA B
SVA C
SVA D
SVA E
SVA F
NES
ESEN / PRO
LSEN / PRO
G
L1HS
Sum Normalized Counts
Truncated
Full Length
Intact
Genic
Intergenic
PRO
ESEN
LSEN
H
L1HS Full Length
5UTR
ORF1
ORF2
3UTR
Read Density (Scaled)
Position (bp)
I
L1HS FL
L1HS_12q13.13_1
L1HS_14q23.2_3
LSEN (Log2 Norm Count)
PRO (Log2 Norm Count)
padj < 0.05
FALSE
TRUE
Genomic Context
Centromeric
CodingTxAdj
Intergenic
Intronic
NoncodingTx
NoncodingTxAdj
J
Gene/TE Expression Correlation
FL
Trnc
Intergenic
Exonic
Intronic
CodingTxAdj
NoncodingTx
NoncodingTxAdj
Rho
(cor)
AluY
HERVK_INT
HERVK_LTR
L1HS
K
Read Density
Norm Read Count (Scaled)
Exonic
Intergenic
Intronic
Repetitive
PRO
ESEN
LSEN
L
TPM Sources
TPM
LowComp
SimpleRep
Retroposon
SAT
DNA
LTR
LINE
SINE
Other
PRO
ESEN
LSEN

Figure 4 | RNA-Seq analysis of TEs. (**A**) Upset plot of differentially expressed repetitive elements, colored by evolutionary age, element length, and element intactness. (**B,C**) Log2 fold changes in DESeq2 normalized expression for repetitive element clades: (**B**) barplot of repeat superfamilies (**C**) heatmap of retrotransposable element families. (**D**) Heatmap of log2 fold changes in DESeq2 normalized expression for selected RTE subfamilies. (**E**) Heatmap of DESeq2 normalized expression of differentially expressed (padj <= 0.05) L1 elements. Elements are grouped by clade and hierarchically clustered by expression. Clade labels are followed by a fraction denoting the number of differentially expressed elements over the total number in the clade. (**F**) Gene set enrichment analysis of selected RTE subfamilies for both the ESEN vs PRO and LSEN vs PRO comparisons. Each RTE subfamily is treated as a gene set. (**G**) Bar plots of total DESeq2 normalized read counts derived from various categories of L1HS elements. (**H**) Composite expression profile of all full-length L1HS elements. (**I**) Scatter plot showing log2 DESeq2 normalized counts of every full length L1HS element, colored by genomic context. Significantly differentially expressed elements are represented by solid circles. CodingTxAdj, coding transcript adjacent (<5 kb); Non-codingTx, embedded in a non-coding transcript; Non-codingTxAdj, non-coding transcript adjacent (<5 kb). (**J**) Heatmap of Spearman correlation coefficients (Rho) for the log2 fold change in expression of RTEs and neighboring genes. The log2 fold change for each RTE was matched to that of the nearest gene. RTEs are grouped as full length or truncated, and then by their genomic context relative to the nearest cellular gene. Significant (p>=0.05) associations are denoted by a *. (**K**) Bar plot of total DESeq2 normalized read counts derived from exonic, intronic, and intergenic compartments, scaled by the total number of counts in each sample**.** (**L**) Bar plot showing the distribution of transcripts per million (TPM) per sample by genomic element class. LowComp, low complexity repeat; SimpleRep,

simple repeat; SAT, satellite sequence; DNA, DNA transposon. Unless otherwise stated, statistical significance was assessed by means of a t-test, with panel-level FDR corrected p-values >= 0.05 *, >=0.01 **, and >=0.001 ***.