Clemson University
TigerPrints

Honors College Theses | Student Works

5-2023

Polyadenylation Mediated by LINE-1

Gillian E. Barnard

Miriam K. Konkel

Follow this and additional works at: https://tigerprints.clemson.edu/hct

Polyadenylation Mediated by LINE-1

Gillian E. Barnard and Miriam K. Konkel

Abstract

Transposable elements (TEs) are sequences that change position within the genome and play an important role in genome expansion. TEs are grouped into two categories based on their transposition mechanism. Class 1 retrotransposons spread via target-primed reverse transcription (RNA to DNA) into different genomic locations. Long interspersed element 1 (L1) is a class 1 retrotransposon that is able to move autonomously, as they encode the protein machinery with an endonuclease and reverse transcriptase activity, to insert themselves back into the genome. L1s were the focus of this study, because they are implicated in creating alternate poly(A) sites in genes. We analyzed 778,128 isoforms produced from 12 samples of long-read RNA (PacBio HiFi) sequencing data to investigate if L1s introduce polyadenylation sites. Isoforms were filtered based on L1 location within the isoforms' 3'UTR, resulting in roughly 3,000 isoforms, spread across 757 genes. L1 subfamilies have arisen throughout evolutionary history due to species-specific substitutions. The L1 subfamilies in the dataset are mostly mammalian specific, while only 43 contain primate specific L1s. The majority of the L1s studied were classified as L1M5 (329), L1ME4b (165), L1MB7 (105), and L1ME4c (105). These L1s contain canonical and noncanonical polyadenylation signals within their 3'UTRs. Alternatively polyadenylated mRNA variants, generated from the same gene, are likely bound by different combinations of *trans*-acting factors that can affect mRNA localization, translation, stability, and decay. Understanding the roles of L1s in alternative polyadenylation will shed light on the impact of TEs on processing efficiency of gene expression.

Introduction

Transposable elements

Transposable elements (TEs) are elements that change position within the genome and are drivers of genome evolution and speciation[36]. More than 50% of the human genome consists of repeated sequences, including interspersed repeats derived from transposable elements[30,51]. Transcripts of repetitive sequences may serve as multifunctional RNAs by participating in the antisense regulation of gene activity and by competing with host-encoded transcripts for cellular factors[43]. After a TE insertion occurs, it is polymorphic in the host population (some have it, some do not). This mutation is subject to population processes of genetic drift (change in frequency of an existing gene variant) and natural selection[33]. The majority of TE insertions are selectively neutral or slightly deleterious, and only a minor fraction, would be of adaptive significance and subjected to positive selection[28]. Some of these deleterious insertions within the host can result in various genetic disorders and cancer; significant association was identified with clinical characteristics of colorectal cancer[7]. TEs are also linked to tumor development in the gastrointestinal tract[7]. TEs contain regulatory sequences: promoters, enhancers, splice sites, and polyadenylation signals, which can actively reshape cellular transcriptomes[19].

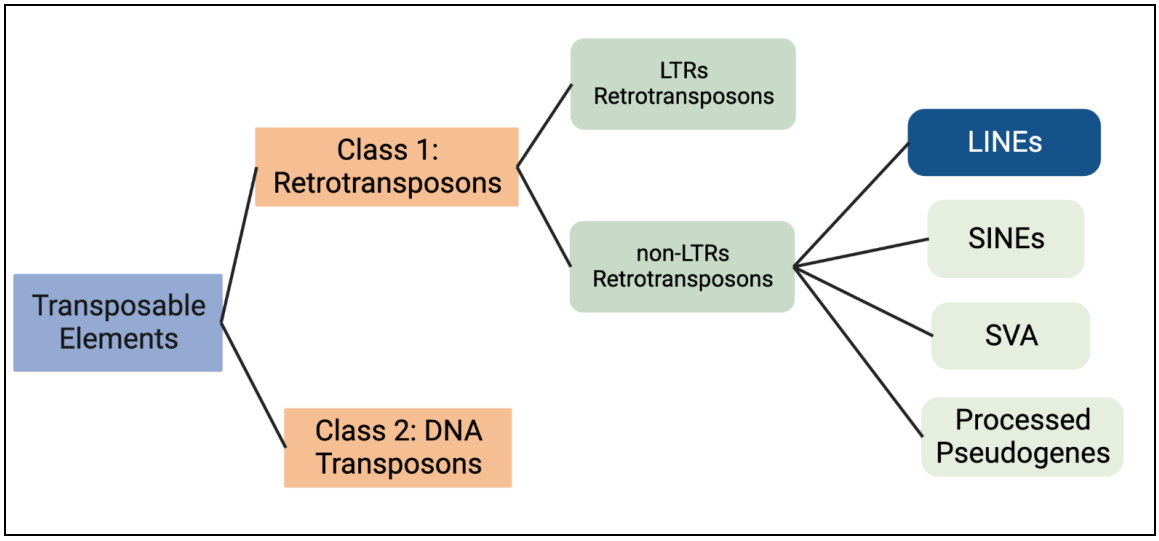


Figure 1. Breakdown of transposable elements. Flow chart shows a breakdown of TEs with specific focus on class 1: retrotransposons. LINEs are a type of class 1 non-LTR retrotransposons.

TEs can be grouped into two categories based on their mechanism of transposition, and further divided into subclasses based on their mechanism of chromosomal integration[6]. The two classes of TEs are: DNA transposons and retrotransposons. DNA transposons are unique from retrotransposons because they have an inverted terminal repeat, encode a transposase, and only constitute about 3% of the human genome[43,51]. DNA transposons mobilize a DNA intermediate, either directly via 'cut and paste' mechanism or a 'peel and paste' replicative mechanism involving a circular DNA intermediate[6]. The transposase mediates the 'cut-and-paste' transposition and, due to this, the insertion site gets duplicated. DNA transposons cannot exercise *cis* preference because their replicative machinery does not allow for it as the transposase has no way to recognize its cognate copy[16]. This is because the transposase cannot identify itself, or distinguish active from inactive elements. As inactive copies (in which the transposase no longer works) accumulate, transposition becomes less efficient. Within the human genome, DNA transposons are considered extinct and no longer move throughout the genome. It is speculated that since DNA transposons utilize a DNA-intermediate, the extinction of DNA transposons is

due to the emergence of host barriers aimed against the cellular entrance of TEs and other forms of invasive DNA[38]. Conversely, retrotransposons comprise about 40% of the human genome due to their retrotransposition mechanism, utilizing an RNA-intermediate[43].

Retrotransposons spread throughout the genome via a copy and paste mechanism. Long interspersed elements (LINEs) are autonomous retrotransposons, as they encode the protein machinery with an endonuclease and reverse transcriptase activity, to insert themselves back into the genome[31]. LINE1 (L1) is a type of LINE clade that accounts for roughly 17% of the human genome and is the main focus of this study[45]. The basic structure of a LINE is composed of: a 5' untranslated region (UTR), an ORF1 (open reading frame 1), an ORF2 (open reading frame 2), and a 3' UTR followed by a polyadenylation signal and poly(A) tail. The proteins expressed are ORF1p and ORF2p and preferentially mobilize encoding RNA in *cis*, but also mobilize *Alu* RNA, and SINE-VNTR-*Alus* (SVAs)[12]. ORF1 encodes an RNA-binding protein (ORF1p) and ORF2 encodes a protein (ORF2p) with an endonuclease (EN) and reverse transcriptase (RT) activity[26]. *Alu* elements require at least ORF2p to mediate their retrotransposition[26].

LINEs

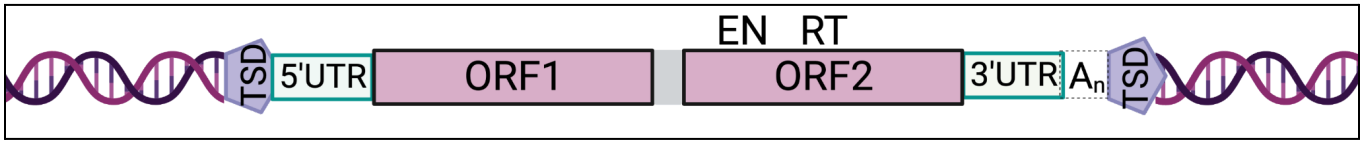


Figure 2. Basic structure of a LINE. A target site duplication (TSD), a 5'UTR, open reading frame 1 (ORF1), ORF2 , a 3'UTR, and a poly(A) signal. ORF1p encodes for an RNA binding protein and ORF2p encodes a protein with endonuclease (EN) and reverse transcriptase (RT) activity.

Full-length LINEs are roughly 6-7 kb in length, but many truncated elements exist within the genome. Roughly 30% of the mammalian genome is composed of LINEs and short interspersed elements (SINEs)[19].

Three distantly related LINE subfamilies ($L1MA_{4\text{-}1}$, $L1PB_{3\text{-}1}$, and $L1PA_{17\text{-}1}$) are found in the human genome. The idea of subfamilies was first suggested after the identification of species-specific substitutions[9]. Most studies point toward the propagation of a single L1 lineage with a linear evolution pattern in mammalian genomes over prolonged periods[21,5,39]. Early in primate evolution roughly three lineages, $L1MA_{4\text{-}1}$, $L1PB_{3\text{-}1}$, and $L1PA_{17\text{-}1}$ were active in parallel for up to 30 million years[22]. The only actively propagating L1 in humans now is L1 *Homo sapiens* (L1Hs)[22]. The human genome contains about 515,000 copies of L1.

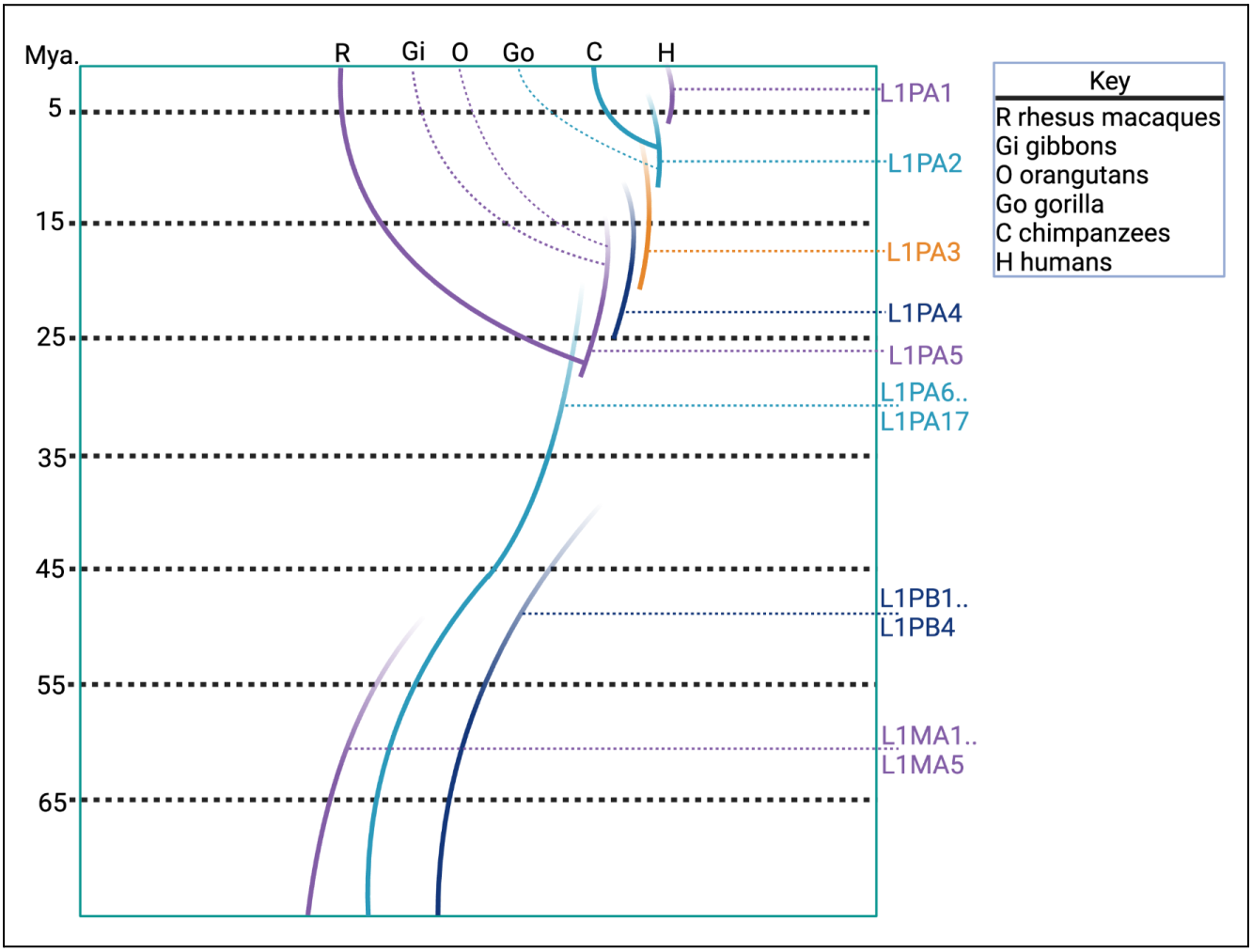


Figure 3. Phylogeny of L1 subfamilies. Phylogeny shows the period of propagating activity of the different L1 subfamilies. Species-specific substitutions branch off to show presence in certain primate-species.

There are some consequences attributed to insertion of TEs in the genome. This should be considered when looking at L1 insertions and their location in a gene. A LINE can be inserted into the promoter of a gene, within an exon, at or near a splice site, or within a non-coding region (intron). When situated within the promoter region, an L1 can alter or disrupt gene expression[47]. When positioned in an exon, it is likely to disrupt the reading frame and result in no gene product[11]. Exonization can also occur when the TEs are incorporated into the transcript and can alter the original gene product[8]. When located at or near the splice site region, it can disrupt the splicing mechanism, alter the reading frame, or result in no protein product[11]. Lastly, if the L1 is localized within a non-coding region (i.e., intron), it can also have negative effects by introducing splice sites and polyadenylation signals[44,37,41]. In general, L1s can target specifically AT-rich regions as insertion sites within the genome[45]. While younger L1 elements are located (on average) closer to genes, full-length elements are more abundant in the sex chromosome than on autosomes[4].

Polyadenylation

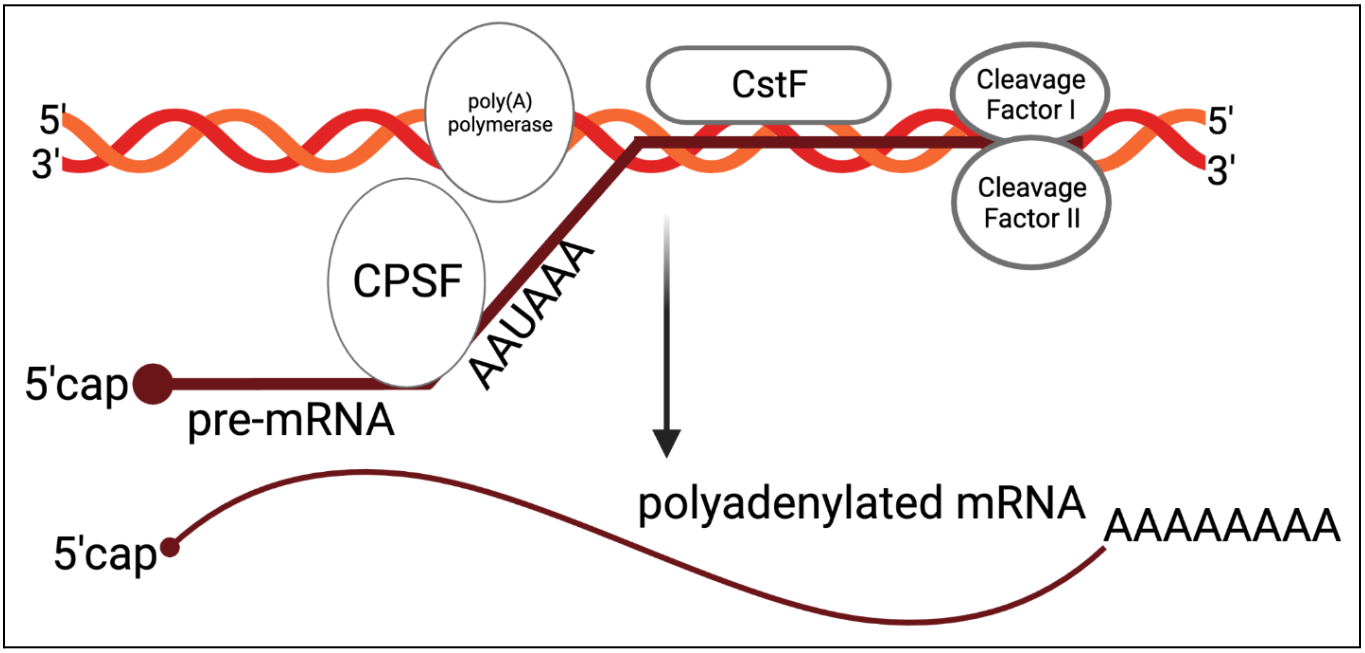


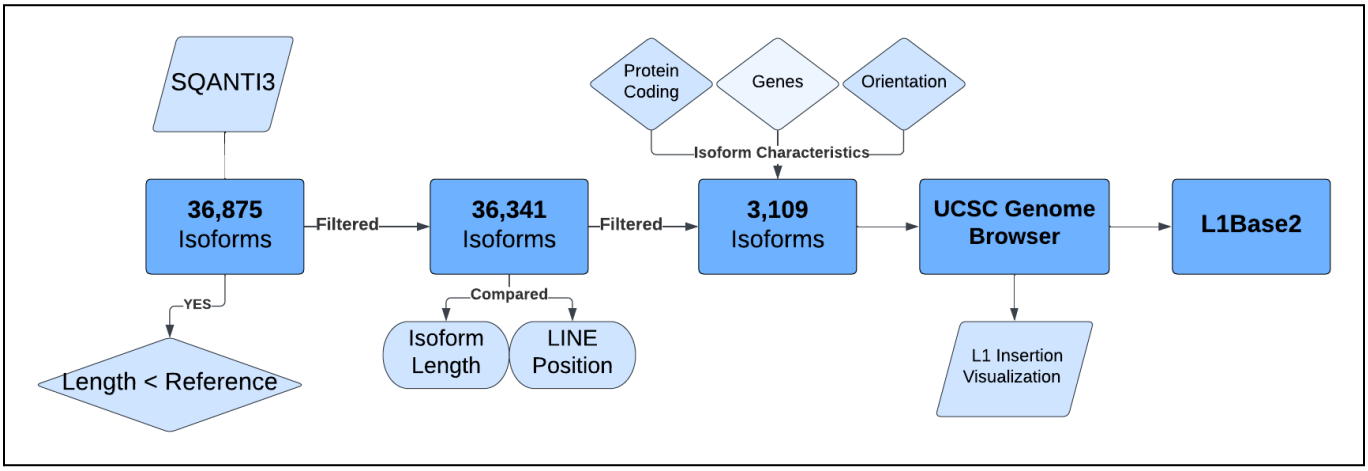
Figure 4. Rendering of a polyadenylation event. Cleavage/polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), a poly(A) polymerase, and cleavage factors I and II.

L1s have been implicated in creating more poly(A) sites when compared to other types of TEs[32]. Polyadenylation is a process that occurs after transcription of a gene (DNA to RNA). A poly(A) signal is found at the 3'end of eukaryotic genes that drives the cleavage and polyadenylation of the pre-mRNA. Cleavage occurs based on a poly(A) signal that is located specifically after a 5'-CA-3' that lies between a canonical AATAAA hexamer (central sequence motif)[42]. This hexamer requires auxiliary elements such as a U- or GU-rich region[42]. Should these polyadenylation signals be located near one another they will compete, often allowing one polyadenylation site to dominate the process[1].

The principal polyadenylation machinery utilizes two cleavage factors, CFI and CFII, the poly(A) polymerase, and two factors involved in RNA sequence recognition: cleavage stimulation factor (CstF) and cleavage/polyadenylation specificity factor (CPSF)[2]. The CstF binds to the downstream GU-rich region, and CPSF binds to the polyadenylation signal[2]. After a cleavage event, the 3'-end [downstream the 5'-CA-3'] is removed, and a poly(A) polymerase adds roughly 200 adenines to the 3'-end of the mRNA. This is to prevent nuclease digestion and to allow the mRNA to successfully leave the nucleus. It also aids the mRNA to be better recognized by the ribosome. The resulting poly(A) tail plays an important role in mRNA translation and stability[48]. A poly(A) binding protein (PABP) becomes bound to the poly(A) tail and the mRNA 5'cap to form a "closed loop" or "circular" mRNA that facilitates translation and protects mRNA from degradation[48]. Alternatively polyadenylated mRNA variants, generated from the same gene, are likely bound by different combinations of *trans*-acting factors that can affect mRNA localization, translation, stability, and decay[48].

TEs can significantly contribute to the creation or modulation of poly(A) sites that are species specific[32]. Some poly(A) sites were encoded by TEs and utilized by endogenous genes, while others are derived from TE regions that have a high propensity to give rise to poly(A) sites[32]. Northern Blot results indicate that premature polyadenylation is conserved in mammalian L1 elements that correlate with high (~40%) A-rich residues in the L1 coding region[41]. L1 elements have devised strategies that allow them to utilize sequences within the 3'UTR and the poly(A) region of L1s to strengthen the usage of their polyadenylation signal[3]. Many transcripts are impacted due to the L1's own internal poly(A) signals, causing premature termination of a gene's transcript. However, these polyadenylation signals are typically relatively weak and a readthrough is commonly expected. The role of L1s in mediating polyadenylation was further investigated in this study. Of the four significant L1 subfamilies observed mediating polyadenylation, two of them were previously recorded as having such roles, while the two remaining had not been previously identified. Understanding the roles of L1s in alternative polyadenylation will shed light on the impact of TEs on processing efficiency of gene expression.

Methodology



The samples used in this study are EBV-transformed lymphocyte cell lines from 12 individuals with diverse genetic backgrounds. They include Finnish, Luhya, Peruvian, Bangladesh, Indian Telugu, Japanese, and Colombian ethnicities. The samples were sequenced by the Human Genome Structural Variation Consortium using PacBio IsoSeq. PacBio IsoSeq is a long-read RNA sequencing approach that generates high-fidelity consensus sequences using SMRTbell (single molecule real-time) technology (0.2% error rate). This data was annotated using SQANTI3 and screened using RepeatMasker[49,46]. SQANTI3 is an automated pipeline that analyzes long-read transcriptomics data[49]. It creates a wide range of summary graphs to aid in the interpretation of the sequencing output, defining up to 47 different descriptors of transcripts and junction properties[49]. RepeatMasker is a software tool widely used in computational genomics to identify, classify, and mask repetitive elements[46].

Shared isoforms can arise across multiple samples. However, this was previously filtered out and the samples were combined. The resulting dataset contained 36,875 unique (no duplications) isoforms with a known L1 present. Therefore, no population comparison analysis could be performed in this study. To identify isoforms with premature polyadenylation events, loci where the length of the isoform was less than the major reference genome length were retrieved as this is an indication of a premature polyadenylation event. The remaining 36,341 isoforms were filtered based on the LINE position within the isoforms. The data was retained if the L1 coordinates were within 0-20 base pairs from the end of the isoform, resulting in 3,109 isoforms.

To confirm if the isoforms containing L1s provided a premature polyadenylation signal and that filtration had been done properly, they were queried against the UCSC Genome Browser using BLAT against the human genome[20]. This process was to visually confirm that the isoforms

contained L1s that were within 20 bp from the end of the isoform. The isoforms were also compared to the GRCh38/hg38 reference transcript to confirm that premature termination had occurred. No such inconsistencies were discovered during this process.

After confirming that the isoform dataset contained L1-mediated polyadenylation loci, the isoform characteristics were then studied for common trends. This included determining if: the isoforms are protein coding, what genes contained these isoforms, and orientation of the L1 within the isoform. The different L1 subfamilies that were observed for each isoform were also studied. Each L1 subfamily found at the end of the isoform was documented in order to determine the L1 subfamilies most commonly found mediating polyadenylation events.

The four most prevalent L1 subfamilies were further studied based on location in the transcript and sequence of their poly(A) signals. The sequences of the isoform were pulled from the SQANTI3 output directory[49]. They were then submitted to the L1Base2 search engine, and the location of the poly(A) signal as well as the sequence was recorded[40]. This was used to compare the rate of canonical (AATAAA) to noncanonical poly(A) signals.

Results

The SQANTI3 annotated data frame was filtered in order to find L1 elements that create an alternate polyadenylation signal in the isoform[49]. There were a total of 3,109 unique (non-shared) isoforms found with the potential for alternate polyadenylation, the majority of which are clustered closer to the end of the isoform. The isoforms were identified in 12 samples of different ethnicities (see Table 1). The final position of the L1 was compared to the length of the isoform and is graphically represented, with most of the isoform L1s observed at the end of the transcript (see Figure 1). Figure 1 is categorized by the base pair difference between the final position of the L1 and the transcript length.

Sample	Ethnicity	Isoform Count
HG00268	Finnish	46
HG01457	Columbian	537
HG02106	Peruvian	106
HG02666	Gambia	58
HG03248	Gambia	1025
HG03807	Bangladesh	70
HG04217	Indian Telugu	885
NA18989	Japanese	55
NA19317	Luhya	107
NA19331	Luhya	88
NA19327	Luhya	65
NA19384	Luhya	67

Table 1: Isoform count for each of the 12 samples. Table shows the sample names used in the study, the ethnicity of that sample, and the isoform count from that sample.

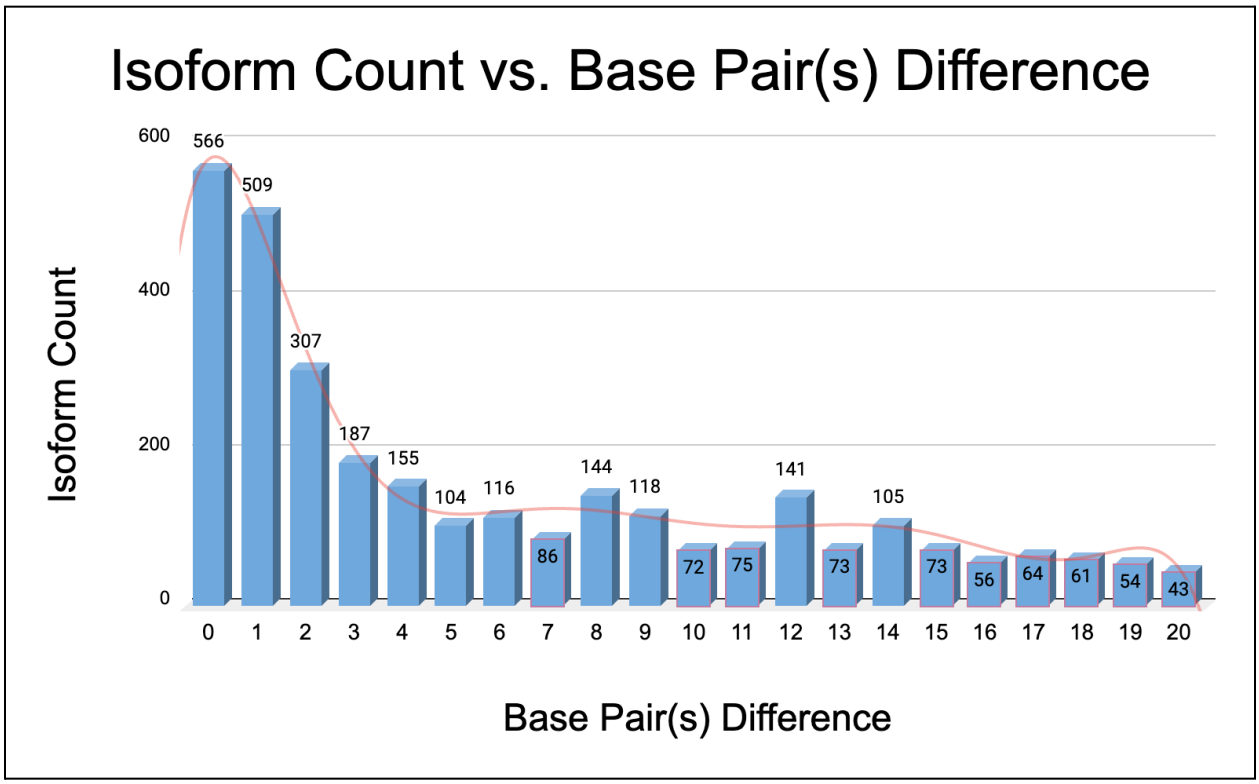


Figure 1. Isoform counts are based on the position of full-length L1s compared to the end of the isoform. Bar graph shows the amount of isoforms with an L1 found across a span of 0 bp to 20 bp from the end of the isoforms.

The majority (2,057 out of 3,109) of the isoforms are protein-coding (see Figure 2). The remaining 1,052 isoforms are noncoding, composed of long-noncoding RNAs and transcribed unprocessed pseudogenes (a DNA sequence that resembles a gene but is inactive)[36]. Across the roughly 3,000 isoforms, 757 genes were identified (see Figure 3). Of these genes, the ones found most often were *OAS3* (56 isoforms) and *PGK1* (49 isoforms). The *OAS3* gene is a part of an enzyme family that plays a role in the inhibition of cellular protein synthesis and viral infection resistance[23]. The *PGK1* gene encodes a glycolytic enzyme but is also known to 'moonlight' as a regulator of metastasis and invasion of hepatocellular carcinoma (HCC) cells[35]. The most common gene family in the dataset is the zinc finger family (79 separate zinc finger genes). Zinc finger proteins harbor diverse functions and include: DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding[27].

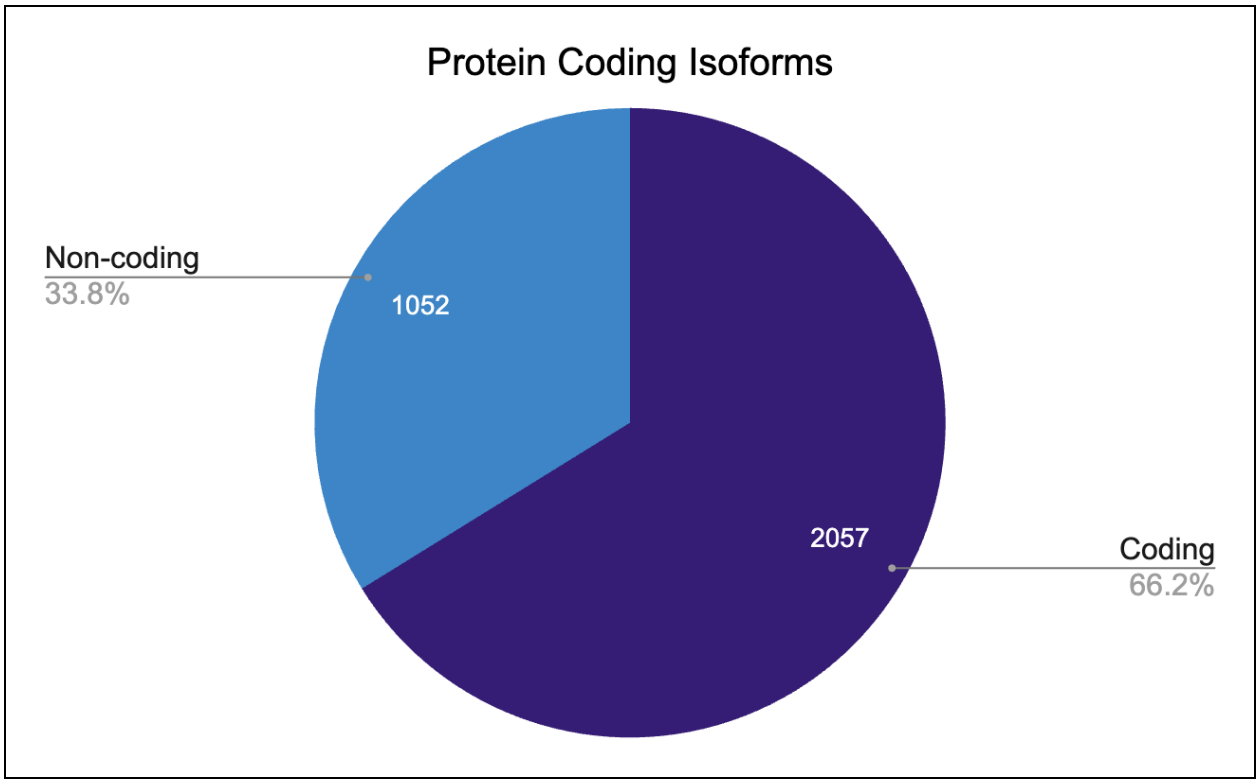


Figure 2. Comparison of protein-coding to non-protein coding isoforms. Percentages of isoforms that are either protein-coding or non-coding. Most of the transcripts are protein-coding (~66%), while the remainder are non-coding (~34%).

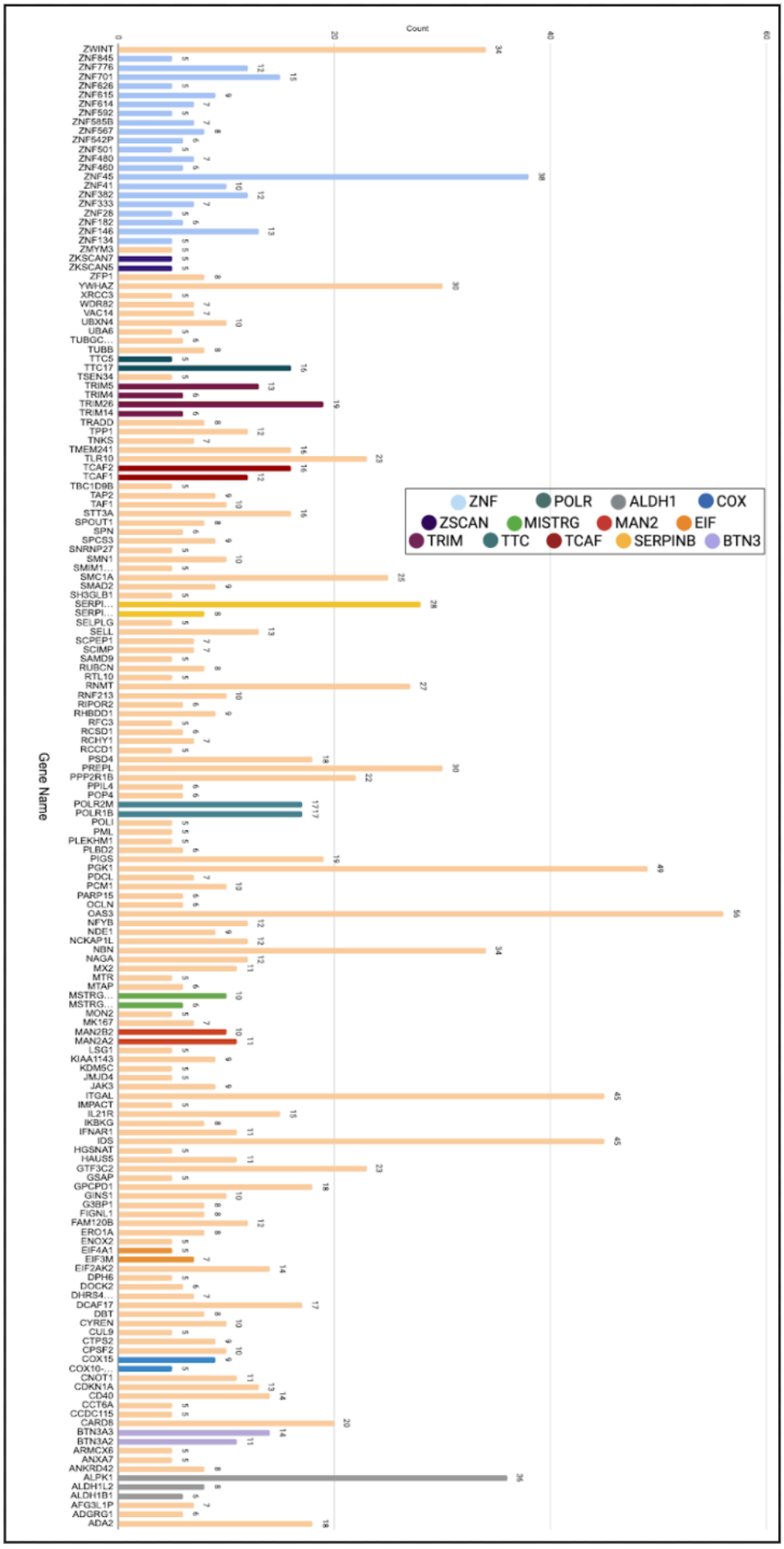

Count
0
20
40
60
Gene Name
ZNF
POLR
ALDH1
COX
ZSCAN
MISTRG
MAN2
EIF
TRIM
TTC
TCAF
SERPINB
BTN3

Figure 3. Genes present within isoform dataset. The graph includes the most common genes found in the isoform dataset. Color-coded and grouped based on gene families present across data.

The isoform dataset was examined for the different L1 subfamilies. This information was retrieved using RepeatMasker[45]. The frequency of the different L1 subfamilies was determined based on their presence in the 3,109 isoforms. These L1s are implicated in creating alternate polyadenylation sites located within a range of 0-20 bp from the end of the transcript. The majority (3,009 out of 3,109) of the L1 subfamilies are mammalian-specific (see Figure 4). Only 100 of the L1s present in the isoforms are classified as non-mammalian specific and are instead primate-specific (43 isoforms) and 'half-L1s' (HAL-1, 57 isoforms). HAL1s are a unique category of L1 elements that encode an ORF1p but not an ORF2p[45]. After documenting the different L1 subfamilies, the strand orientation was also recorded (see Figure 5) as L1's can be integrated in a sense or antisense orientation. The majority of the L1s were in a preferential antisense orientation (1,814), while the rest of the L1s (1,295) were integrated in the sense orientation. The L1M5 (329), L1ME4b (165), L1MB7 (105), and L1ME4c (105) subfamilies are indicated as creating the most abundant premature poly(A) sites among the isoforms (see Figure 4).

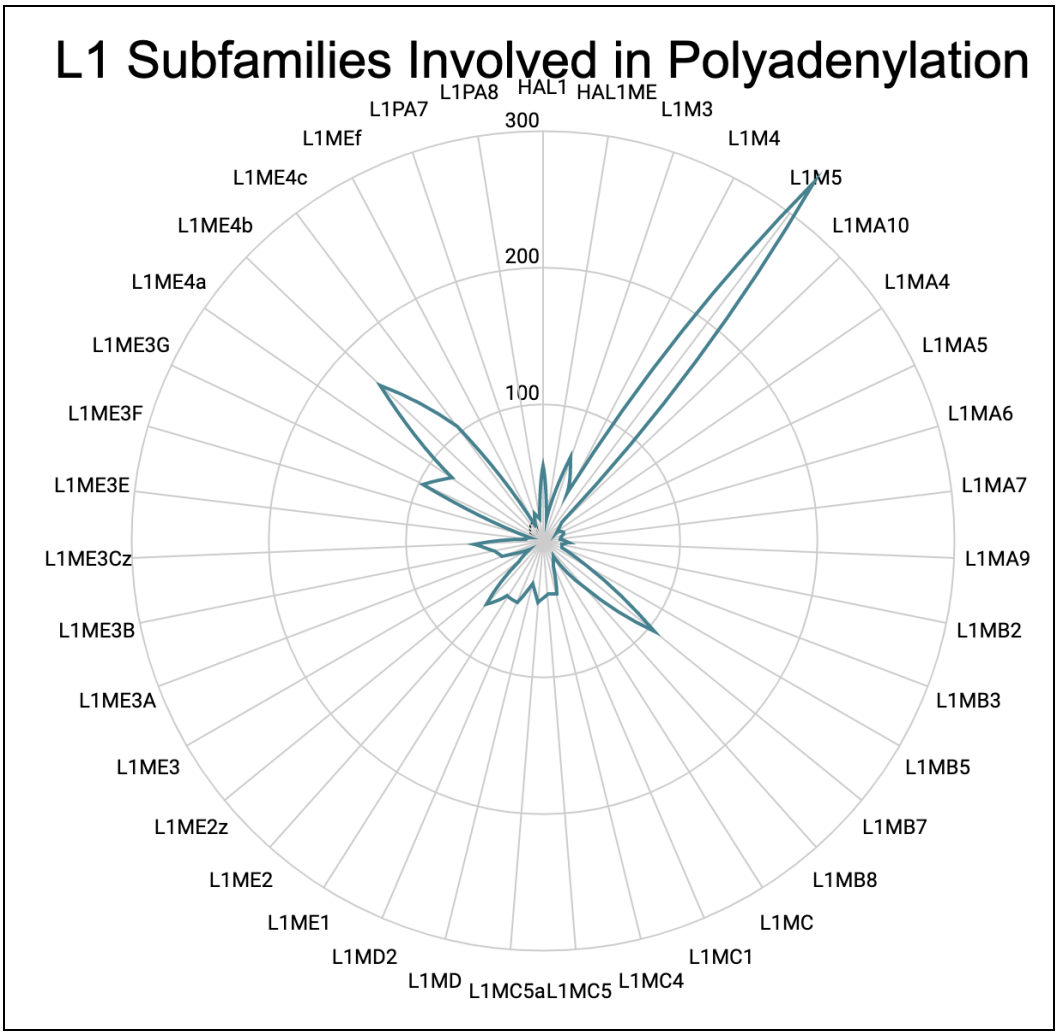


Figure 4. L1 Prevalence in isoform polyadenylation sites. Radar graph shows all the different L1 subfamilies found across the isoforms. Prevalence and amount is indicated by length of line.

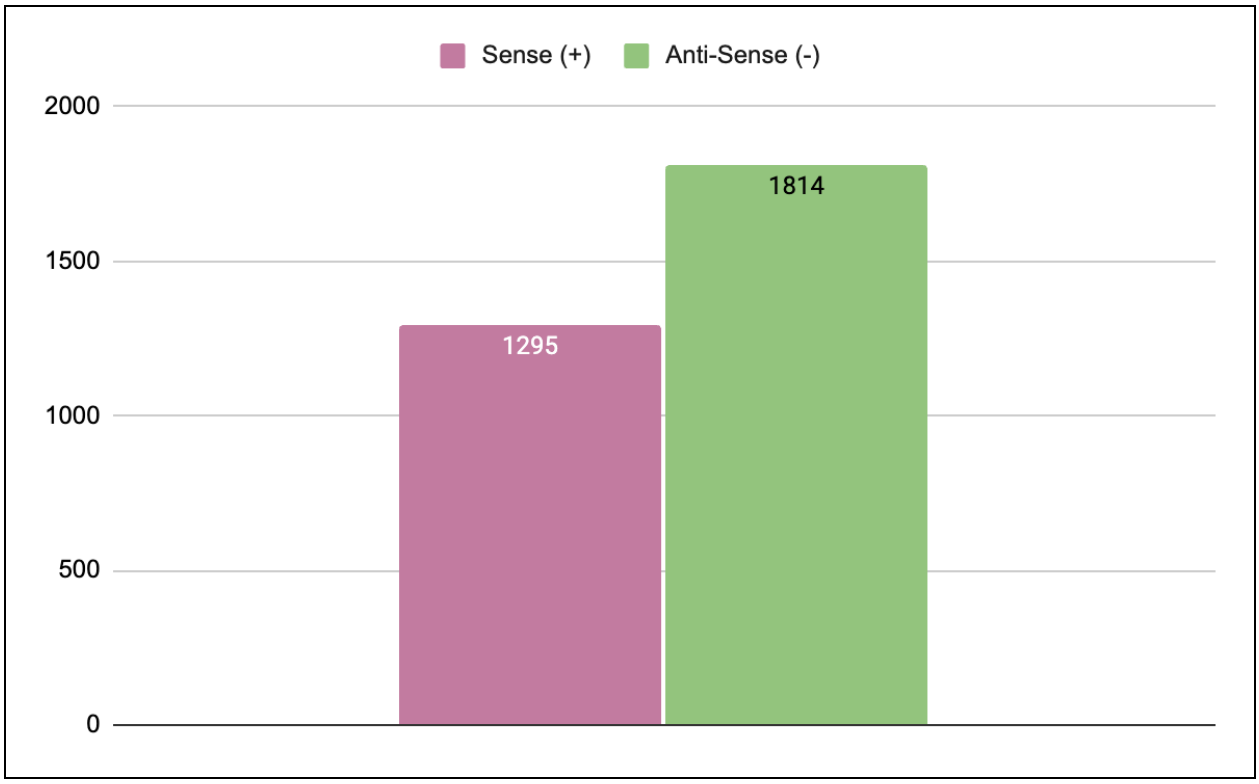


Figure 5. L1 strand orientation within isoforms. Bar graph showing the amount of L1s orientation within the transcript. Most of the L1s are in the antisense orientation (1,814 out of 3,109).

The L1M5 subfamily was identified across different types of genes (see Figure 6). Of the 329 isoforms with an L1M5, 79 genes were found. L1M5 was found in the sense orientation in only two of these genes (*POLR1D* and *NOL9*). Each isoform sequence that contained an L1M5 was retrieved and analyzed with the L1Base2 tool[40]. L1Base2 is a dedicated database containing putatively active L1 insertions residing in humans and was used to determine L1 poly(A) signals within transcripts[40]. The canonical (AATAAA) and noncanonical poly(A) signals were recorded in the L1M5 subfamily (see Table 2). A noncanonical poly(A) signal must have at least one or more changes in the hexameric sequence[42]. The majority of the poly(A) signals were canonical (76%). The next most common signal was a single base change at site 4 (A -> T) (16%), then a base change at site 6 (A -> T) (6%), lastly a base change at site 3 (A -> C) (2%). All of the signals were found within the 3'UTR of the L1.

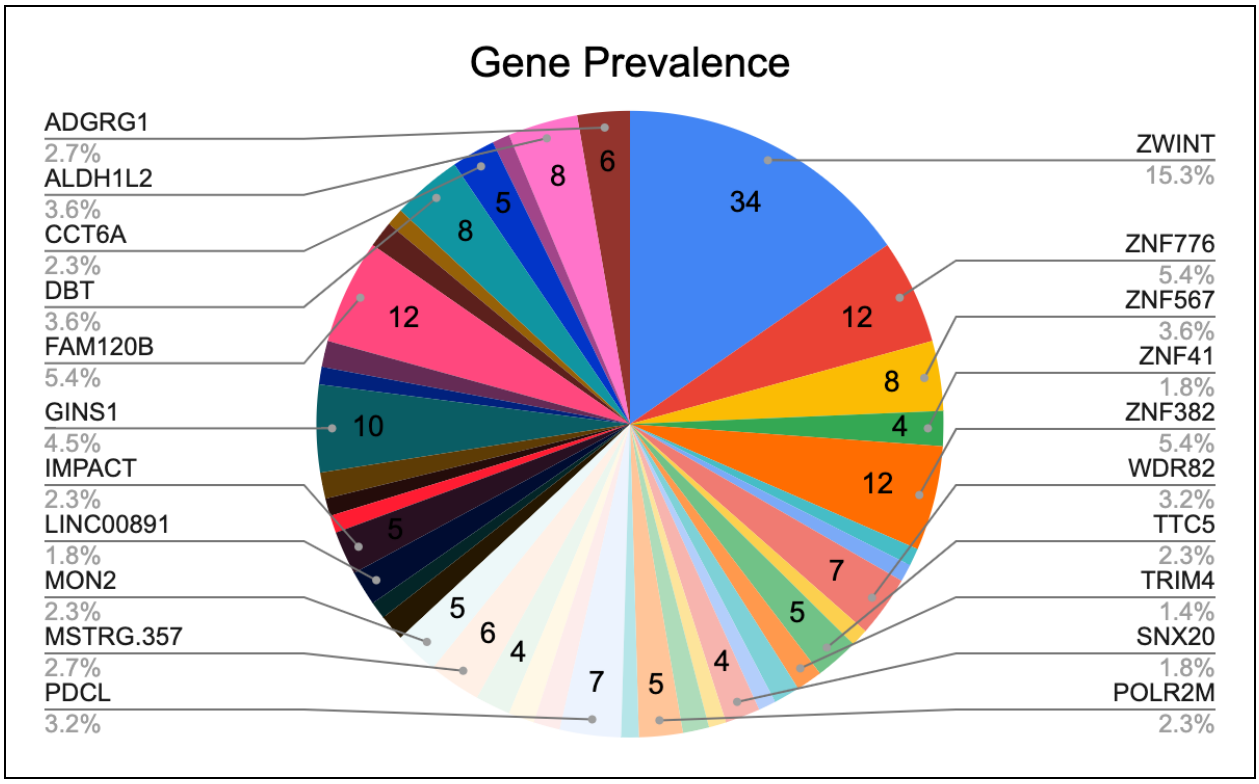


Figure 6. Percentage of genes with an L1M5 insertion. Most common genes with an L1M5 mediating polyadenylation events. Most common gene was *ZWINT* (34 out of 329 L1M5 isoforms).

1	2	3	4	5	6	
A	**A**	**T**	**A**	**A**	**A**	76%
A	**A**	**T**	**T**	**A**	**A**	16%
A	**A**	**T**	**A**	**A**	**T**	6%
A	**A**	**C**	**A**	**A**	**A**	2%

Table 2. L1M5 poly(A) signal(s) found within the 3'UTR. Data came from L1Base2 analysis using transcript sequence[40]. Yellow highlighted sequence is the canonical poly(A) signal. Percentages indicated how often the signals were found across the 329 isoforms. Three noncanonical pol(A) signals were observed.

The L1ME4b subfamily was observed across different types of genes (see Figure 7). Of the 165 isoforms with an L1ME4b, 36 genes were found. Of the 36 genes, 10 genes (*ZNF182, TLR10, SCAF4, RAP2C-AS1, POLA2, PDRG1, LAMTOR3, CCDC93, ARHGAP31,* and *ABCB8*)

were found with L1ME4b in the sense orientation. Each isoform sequence that contained L1ME4b was retrieved and analyzed with the L1Base2 tool (see Table 3)[40]. The majority of the poly(A) signals were canonical (62%). The next most common signal was a base change at site 4 (A -> T) (31%); lastly, a base change at site 6 (A -> T) (7%).

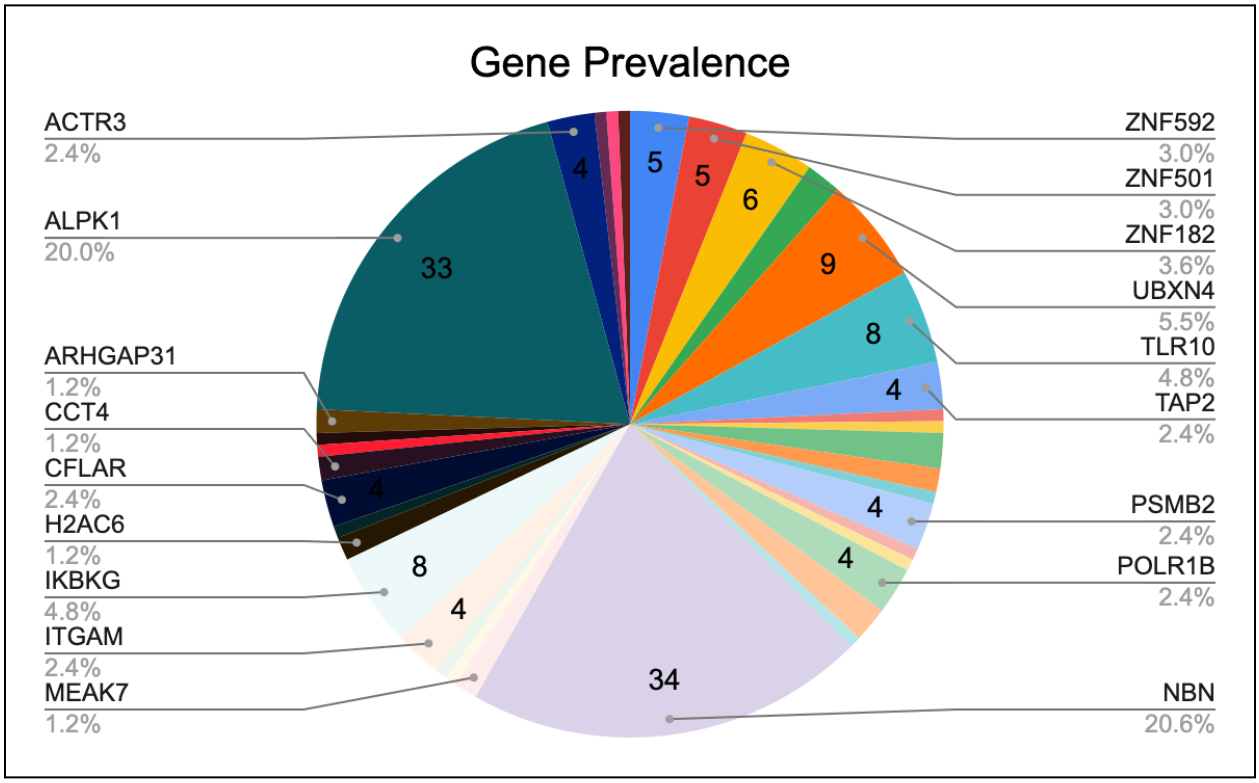


Figure 7. Percentage of genes within an L1ME4b insertion. Most common genes with an L1ME4b mediating polyadenylation signal. Most common gene was *NBN* (34 out of 165 L1ME4b isoforms).

1	2	3	4	5	6	
A	**A**	**T**	**A**	**A**	**A**	62%
A	**A**	**T**	**T**	**A**	**A**	31%
A	**A**	**T**	**A**	**A**	**T**	7%

Table 3. L1ME4b poly(A) signal(s) found within the 3'UTR. Data came from L1Base2 analysis using transcript sequence[40]. Yellow highlighted sequence is the canonical poly(A) signal. Percentages indicated how often the signals were found across the 165 isoforms. Only two noncanonical poly(A) signals were discovered.

The L1MB7 subfamily was recognized across different types of genes (see Figure 8). 20 genes were identified across the 105 isoforms. Six of these genes (*OXTR, MAN2B2, KIAA0513, ALDH1B1, AFG3L1P* and *AC018445.5*) are found with L1MB7 in the sense orientation. The canonical and noncanonical poly(A) signals were recorded in the L1MB7 subfamily (see Table 4). The majority of the poly(A) signals were noncanonical (54%), with a base change at site 6 (A -> T). The second most common signal was canonical (33%). The least common signal contained a base change at site 4 (A -> T) (13%).

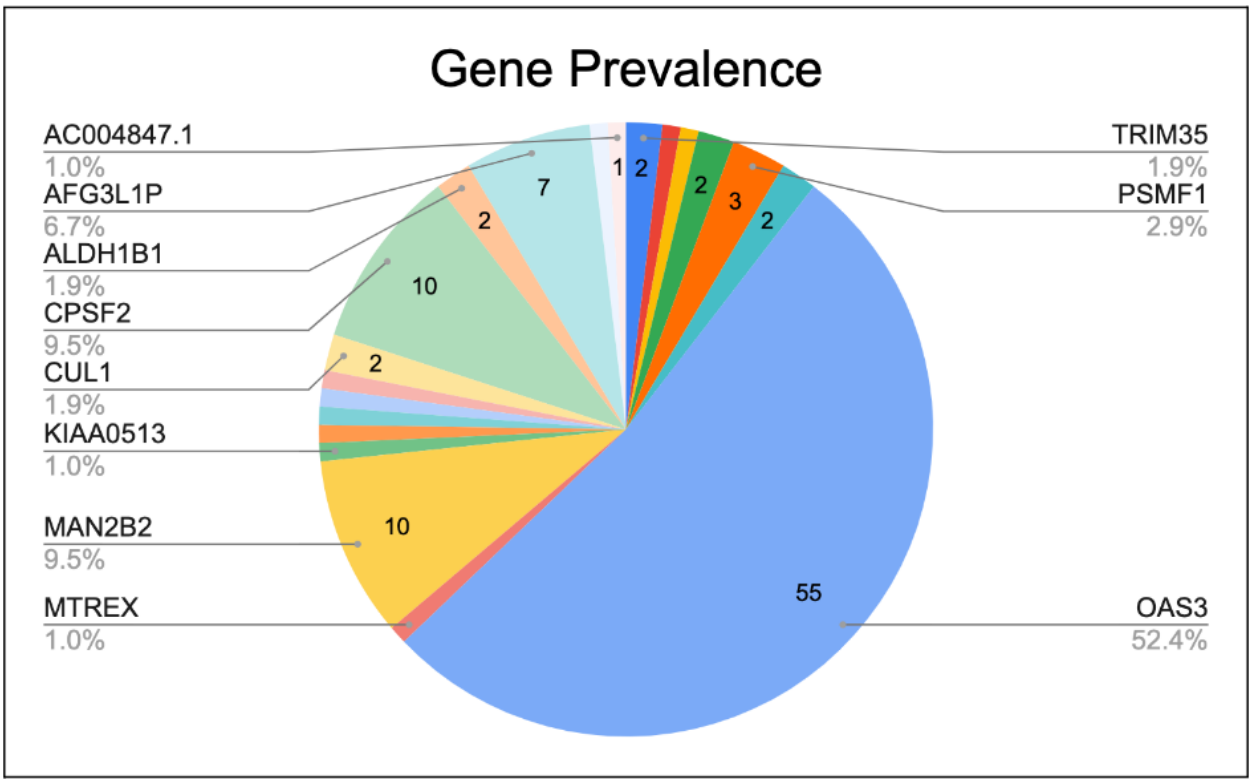


Figure 8. Percentage of genes with an L1MB7 insertion. Graph demonstrates common genes with an L1MB7 mediating polyadenylation signal. Most common gene was *OAS3* (55 out of 105 L1MB7 isoforms).

1	2	3	4	5	6	
A	A	T	A	A	A	33%
A	A	T	T	A	A	13%
A	A	T	A	A	T	54%

Table 4. L1MB7 poly(A) signal(s) found within the 3'UTR. Data came from L1Base2 analysis using transcript sequence[40]. Yellow highlighted sequence is the canonical poly(A) signal. Percentages indicated how often the signals were found across the 105 isoforms. Only two noncanonical poly(A) signals were discovered.

The L1ME4c subfamily was linked to different types of genes (see Figure 9). There are 105 isoforms with an L1ME4c present in the 3'UTR and only 20 genes observed. Within five of these genes (*ZNF606, PUS7L, MAML1, FCGR2A* and *DNAJC7*) L1ME4c is in the sense orientation. After using the L1Base2, the canonical (AATAAA) and noncanonical poly(A) signals were recorded in the L1ME4c subfamily (see Table 5)[40]. The majority of the poly(A) signals were canonical (68%). The only other signal was a base change at site 4 (A -> T) (32%).

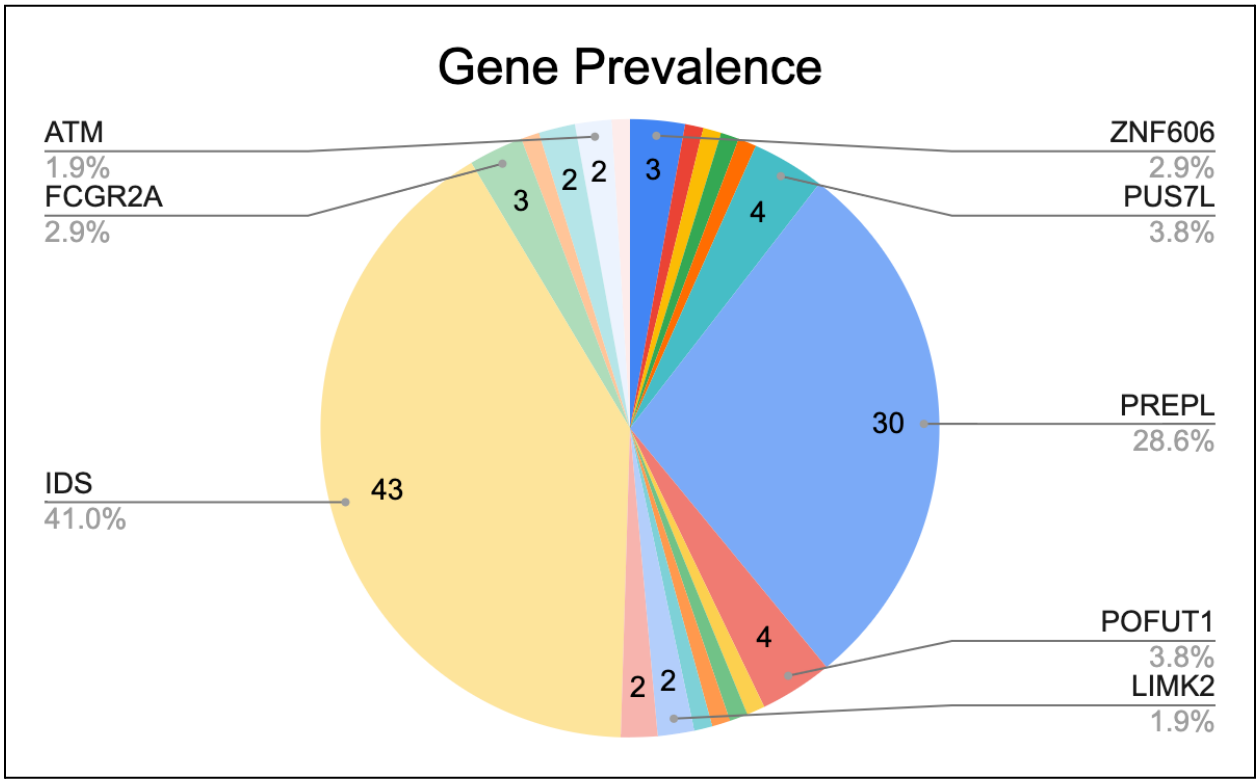


Figure 9. Percentage of genes with an L1ME4c insertion. Graph demonstrates common genes

with an L1ME4c mediating polyadenylation signal. The most common gene was *IDS* (43 out of 105 L1ME4c isoforms).

1	2	3	4	5	6	
A	**A**	**T**	**A**	**A**	**A**	68%
A	**A**	**T**	**T**	**A**	**A**	32%

Table 5. L1ME4c poly(A) signal(s) found within the 3'UTR. Data came from L1Base2 analysis using transcript sequence[40]. Yellow highlighted sequence is the canonical poly(A) signal. Percentages indicated how often the signals were found across the 105 isoforms. Only one noncanonical poly(A)signal was discovered.

Discussion

Polyadenylation is the process by which adding roughly 200 adenines to the 3'end of an mRNA. The 3' end of a transcript plays an important role in the development of pre-mRNA to mature mRNA, with the largest weight on the 3' poly(A) tail. 3,109 isoforms found in the study contain an L1 sequence within the 3'UTR, introducing alternate poly(A) signals and mediating the polyadenylation process. The premature transcripts discovered have an L1 within it that has utilized sequences within the 3'UTR and the poly(A) region to strengthen the usage of their polyadenylation signal[3].

The majority of human genes (>80 %) yield multiple mRNA isoforms with alternative 3′UTRs due to differences in the position of 3′ end cleavage and polyadenylation[25]. With an L1 within the 3'UTR of a transcript, it can introduce a premature polyadenylation event. Since posttranscriptional regulatory sequences are contained within 3′UTRs, alternatively polyadenylated mRNA variants generated from the same gene are likely to be bound by different combinations of *trans*-acting factors (proteins and microRNAs)[48]. MicroRNAs (miRNA) are small, single-stranded, non-coding RNA molecules that are involved in RNA silencing and post-transcriptional regulation of gene expression[24]. Changes in the position of cleavage and

initiation of polyadenylation have the potential to impact downstream events in mRNA.

The majority of the L1 subfamilies are mammalian-specific (only 100 isoforms are not confirmed to contain mammalian-specific L1s). Mammalian-specific L1s were last actively propagating over 65 million years ago, meaning the majority of these L1s are truncated and/or have accumulated random mutations[22]. The four most prevalent L1 subfamilies examined in this study (L1M5, L1ME4b, L1MB7, and L1ME4c) contained both canonical and noncanonical poly(A) signals. A canonical poly(A) signal is an AATAAA hexameric sequence. Mutations can accumulate within the poly(A) signal (in the consensus sequence); positions 1, 2, and 5 are tolerant to point mutations, while positions 3, 4, and 6 are highly conserved[2]. Interestingly, the mutations within the L1s variant poly(A) signals are located at sites 3, 4, and 6. This is further confirmed by the fact that these mutations are not found in the consensus sequence.

As the number of poly(A) sites in an mRNA molecule increases, the proportion of canonical AATAAA signals decreases[2]. In instances where mRNAs have multiple poly(A) sites along the length of the transcript, they tend to use a higher proportion of noncanonical signals. The basis for the occurrence of variant polyadenylation signals is that variation of control sequences mediates variation in polyadenylation rates, thus regulating gene expression[13]. For instance, in mRNAs that contain both a canonical and a noncanonical signal in their 3'UTR, the noncanonical signal is processed less efficiently[2]. Meaning, the mutations accumulated within an evolutionarily old L1, within its poly(A) site, can introduce noncanonical poly(A) signals in transcripts that are processed less efficiently and drive non-major isoforms.

Conclusion

L1s are able to mediate polyadenylation events within EBV-transformed lymphocytes. The samples used for this study were from 12 individuals and 3,109 isoforms. The most common

L1 subfamilies found were mammalian-specific. These subfamilies are very old and are truncated with mutations. This likely allowed for the introduction of both canonical and noncanonical poly(A) signals. These noncanonical signals may influence downstream effects on mRNA stability, translation efficiency, or localization of an mRNA in a tissue-specific manner. Understanding the roles of L1s in alternative polyadenylation will shed light on the impact of TEs on processing efficiency of gene expression.

Future Directions

There are known downstream effects of noncanonical poly(A) signals on mRNA localization, translation, stability, and decay. Future studies will assess the processing efficiency of those polyadenylation signals in relation to their sequence and position in the 3'UTR.

Acknowledgments

This work is supported in part by the "Center of Biomedical Research Excellence (COBRE) in Human Genetics" 1P20GM139769, Clemson Creative Inquiry Program, and the National Institute of General Medical Sciences (NIGMS). Clemson University is acknowledged for the Palmetto cluster resource.

Special Regards

I would also like to acknowledge the members of the CI Tangible Genomics: **Dr. Miriam K. Konkel**, Ashley Kirby, **Mark Loftus**, **Ashtyn Hill**, Gianni Martino, Kassandra Roemer, Olivia Duffy, and Emily Golba.

Common Abbreviations

TEs	Transposable Elements
LINEs	Long Interspersed Elements
L1	Long Interspersed Element-1
SINEs	Short Interspersed Elements

UTR	Untranslated region
L1Hs	L1 *Homo sapiens*
SVA	SINE-VNTR-*Alus*
CstF	Cleavage Stimulation Factor
CPSF	Cleavage and Polyadenylation Specific Factor
SMRT	Single molecule real-time

References

1. Batt DB, Luo Y, Carmichael GG, Polyadenylation and transcription termination in gene constructs containing multiple tandem polyadenylation signals, *Nucleic Acids Research*, Volume 22, Issue 14, 25 July 1994, Pages 2811–2816 https://doi.org/10.1093/nar/22.14.2811
2. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 2000 Jul;10(7):1001-10. doi: 10.1101/gr.10.7.1001. PMID: 10899149; PMCID: PMC310884.
3. Belancio VP, Whelton M, Deininger P, Requirements for polyadenylation at the 3′ end of LINE-1 elements, Gene, Volume 390, Issues 1–2, 2007, Pages 98-107, ISSN 0378-1119, https://doi.org/10.1016/j.gene.2006.07.029.
4. Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol. 2001 Jun;18(6):926-35. doi: 10.1093/oxfordjournals.molbev.a003893. PMID: 11371580.
5. Boissinot S, Entezam A, Young L, Munson PJ, Furano AV. 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res 14:1221–1231.
6. Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1577-z
7. Cajuso, T., Sulo, P., Tanskanen, T. *et al.* Retrotransposon insertions can initiate colorectal cancer and are associated with poor survival. *Nat Commun* 10, 4022 (2019). https://doi.org/10.1038/s41467-019-11770-0
8. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009 Oct;10(10):691-703. doi: 10.1038/nrg2640. PMID: 19763152; PMCID: PMC2884099.
9. Daniels GR, Fox GM, Loewensteiner D, Schmid CW, Deininger PL. 1983. Species-specific homogeneity of the primate Alu family of repeated DNA sequences. Nucleic Acids Res 11:7579–7593.
10. Dazenière, J., Bousios, A., & Eyre-Walker, A. (2022). Patterns of selection in the evolution

of a transposable element. *G3 (Bethesda, Md.)*, *12*(5). https://doi.org/10.1093/g3journal/jkac056
11. Deininger, Prescott L. John V Moran, Mark A Batzer, Haig H Kazazian, Mobile elements and mammalian genome evolution, Current Opinion in Genetics & Development, Volume 13, Issue 6, 2003, https://doi.org/10.1016/j.gde.2003.10.013.
12. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y., & Moran, J. v. (2015). A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Molecular Cell*, *60*(5), 728–741. https://doi.org/10.1016/j.molcel.2015.10.012
13. Edwalds-Gilbert G et al., Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res. 1997 Jul 1;25(13):2547-61. doi: 10.1093/nar/25.13.2547. PMID: 9185563; PMCID: PMC146782.
14. Graham T, Boissinot S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. J Biomed Biotechnol. 2006;2006(1):75327. doi: 10.1155/JBB/2006/75327. PMID: 16877820; PMCID: PMC1510949.
15. Han, J. S., Szak, S. T., & Boeke, J. D. (2004). *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes*. www.nature.com/nature
16. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA. L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A. 2008 Dec 9;105(49):19366-71. doi: 10.1073/pnas.0807866105. Epub 2008 Nov 26. PMID: 19036926; PMCID: PMC2614767.
17. Hancks, D. C., & Kazazian, H. H. (2012). Active human retrotransposons: Variation and disease. In *Current Opinion in Genetics and Development* (Vol. 22, Issue 3, pp. 191–203). https://doi.org/10.1016/j.gde.2012.02.006
18. Huang, W., Tsai, L., Li, Y., Hua, N., Sun, C., & Wei, C. (2017). Widespread of horizontal gene transfer in the human genome. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3649-y
19. Kazazian HH, Goodier JL, Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites, Cell, 2008, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2008.09.022.
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.
21. Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16:78–87.
22. Konkel MK, Walker JA, Batzer MA. LINEs and SINEs of primate evolution. Evol Anthropol. 2010 Nov 1;19(6):236-249. doi: 10.1002/evan.20283. PMID: 25147443; PMCID: PMC4137791.
23. Kristiansen H, Gad HH, Eskildsen-Larsen S, Despres P, Hartmann R. The oligoadenylate synthetase family: an ancient protein family with multiple antiviral activities. J Interferon Cytokine Res. 2011 Jan;31(1):41-7. doi: 10.1089/jir.2010.0107. Epub 2010 Dec 12. PMID: 21142819.
24. Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M. Silencing of microRNAs in vivo with 'antagomirs'. Nature. 2005 Dec 1;438(7068):685-9. doi: 10.1038/nature04303. Epub 2005 Oct 30. PMID: 16258535.
25. Kühn U et al., Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J Biol Chem. 2009 Aug 21;284(34):22803-14. doi:

10.1074/jbc.M109.018226. Epub 2009 Jun 9. PMID: 19509282; PMCID: PMC2755688.
26. Kulpa, D. A., & Moran, J. v. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nature Structural and Molecular Biology*, *13*(7), 655–660. https://doi.org/10.1038/nsmb1107
27. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 2001 Feb;11(1):39-46. doi: 10.1016/s0959-440x(00)00167-6. PMID: 11179890.
28. Lambert ME, McDonald JF, Weinstein IB.. 1988. *Eukaryotic transposable elements as mutagenic agents*. Cold Spring Harbor (NY:): Cold Spring Harbor Laboratory Press.
29. Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. In *Nature Reviews Genetics* (Vol. 21, Issue 12, pp. 721–736). Nature Research. https://doi.org/10.1038/s41576-020-0251-y
30. Lander, S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Yeh, R.-F. (2001). Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center. In *NATURE* (Vol. 409). www.nature.com
31. Lavie, L., Maldener, E., Brouha, B., Meese, E. U., & Mayer, J. (2004). The human L1 promoter: Variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Research*, *14*(11), 2253–2260. https://doi.org/10.1101/gr.2745804
32. Lee, J. Y., Ji, Z., & Tian, B. (2008). Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Research*, *36*(17), 5581–5590. https://doi.org/10.1093/nar/gkn540
33. Le Rouzic A, Boutin TS, Capy P. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 2007 Dec 4;104(49):19375-80. doi: 10.1073/pnas.0705238104. Epub 2007 Nov 26. PMID: 18040048; PMCID: PMC2148297.
34. Liu, M., & Eiden, M. v. (2011). Role of human endogenous retroviral long terminal repeats (LTRs) in maintaining the integrity of the human germ line. *Viruses*, *3*(6), 901–905. https://doi.org/10.3390/v3060901
35. McCarrey, J., Thomas, K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* 326, 501–505 (1987). https://doi.org/10.1038/326501a0
36. McClintock B. The Order of the Genes C, Sh and Wx in Zea Mays with Reference to a Cytologically Known Point in the Chromosome. Proc Natl Acad Sci U S A. 1931 Aug;17(8):485-91. doi: 10.1073/pnas.17.8.485. PMID: 16587653; PMCID: PMC1076097.
37. Meischl C, Boer M, Ahlin A, Roos D. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. Eur J Hum Genet. 2000 Sep;8(9):697-703. doi: 10.1038/sj.ejhg.5200523. PMID: 10980575.
38. Pace JK 2nd, Feschotte C. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 2007 Apr;17(4):422-32. doi: 10.1101/gr.5826307. Epub 2007 Mar 5. PMID: 17339369; PMCID: PMC1832089.
39. Pascale E, Valle E, Furano AV. 1990. Amplification of an ancestral mammalian L1 family of

long interspersed repeated DNA occurred just before the murine radiation. Proc Natl Acad Sci USA 87:9481–9485.
40. Penzkofer T, et al., L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. Nucleic Acids Res. 2017 Jan 4;45(D1):D68-D73. doi: 10.1093/nar/gkw925. Epub 2016 Oct 18. PMID: 27924012; PMCID: PMC5210629.
41. Perepelitsa-Belancio, V., & Deininger, P. (2003). RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genetics*, *35*(4), 363–366. https://doi.org/10.1038/ng1269
42. Proudfoot, N. J. (2011). Ending the message: Poly(A) signals then and now. In *Genes and Development* (Vol. 25, Issue 17, pp. 1770–1782). https://doi.org/10.1101/gad.17268411
43. Schumann, G. G., Gogvadze, E. v, Osanai-Futahashi, M., Kuroki, A., Münk, C., Fujiwara, H., Ivics, Z., & Buzdin, A. A. (n.d.). *Unique Functions of Repetitive Transcriptomes*. https://doi.org/10.1016/S1937-6448(10)85003-8
44. Schwahn U, Lenzner S, Dong J, et al. Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nature Genetics. 1998;19(4):327–332.
45. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev.* 1999;9:657–63.
46. Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*.2008-2015 <http://www.repeatmasker.org>.
47. Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol. 2001 Mar;21(6):1973-85. doi: 10.1128/MCB.21.6.1973-1985.2001. PMID: 11238933; PMCID: PMC86790.
48. Sweet TJ, Licatalosi DD. 3' end formation and regulation of eukaryotic mRNAs. Methods Mol Biol. 2014;1125:3-12. doi: 10.1007/978-1-62703-971-0_1. PMID: 24590775; PMCID: PMC6872190.
49. Tardaguila M, de la Fuente L, Marti C, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*, 2018. 28(3):396-411. doi:10.1101/gr.222976.117
50. Wheelan, S. J., Aizawa, Y., Han, J. S., & Boeke, J. D. (2005). Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Research*, *15*(8), 1073–1078. https://doi.org/10.1101/gr.3688905
51. Wheeler DA et al., The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002;3(12):RESEARCH0084. doi: 10.1186/gb-2002-3-12-research0084. Epub 2002 Dec 23. PMID: 12537573; PMCID: PMC151186.