 doi:10.20944/preprints202011.0548.v1

Role of transposable elements in gene regulation in the human genome

Arsala Ali[1], Kyudong Han[2,3], Ping Liang[1,4*]

[1] Department of Biological Sciences, Brock University, St. Catharines, Ontario, Canada L2S 3A1

[2] Department of Microbiology & [3] Center for Bio-Medical Engineering Core Facility, Dankook University, 119 Dandae-ro, Dongnam-gu, Cheonan, Chungnam 31116, Korea

[4] Centre of Biotechnologies, Brock University, St. Catharines, Ontario, Canada L2S 3A1

[*] Corresponding author: Ping Liang, Department of Biological Sciences, Brock University, St. Catharines, Ontario, Canada; Tel: 1-905-688-5550 ext. 5922; fax: 1-905-688-1855; email: pliang@brocku.ca

Abstract: Transposable elements (TEs), also known as mobile elements (MEs), are interspersed repeats that constitute a major fraction of the genomes of higher organisms. As one of their important functional impacts on gene function and genome evolution, TEs participate in regulating the expression of genes nearby and even far away at transcriptional and post-transcriptional levels. There are two principal ways by which TEs regulate expression of genes in the human genome. First, TEs provide *cis*-regulatory sequences in the genome. TEs' intrinsic regulatory properties for their own expression make them potential factors for regulating the expression of the host genes. TE-derived *cis*-regulatory sites are found in promoter and enhancer elements, providing binding sites for a wide range of trans-acting factors. Second, TEs encode for regulatory RNAs. TEs sequences have been revealed to be present in a substantial fraction of miRNAs and long non-coding RNAs (lncRNAs), indicating their TE origin. Furthermore, TEs sequences were found to be critical for regulatory functions of these RNAs including binding to the target mRNA. TEs thus provide crucial regulatory roles by being part of *cis*-regulatory and regulatory RNA sequences. Moreover, both TE-derived *cis*-regulatory sequences and TE-derived regulatory RNAs, have been implicated to provide evolutionary novelty to gene regulation. These TE-derived regulatory mechanisms also tend to function in tissue-specific fashion. In this review, we aim to comprehensively cover the studies regarding these two aspects of TE-mediated gene regulation, mainly focusing on the mechanisms, contribution of different types of TEs, differential roles among tissue types, and lineage specificity, based on data mostly in humans.

Keywords: transposable elements, mobile elements, gene regulation, evolution, human

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

1. Overview of transposable elements and their role in the human genome

Transposable elements (TEs), also known as mobile elements (MEs), are interspersed repeats constituting a major fraction of the genomes in higher organisms. The contribution of TEs in the human genome has been updated to 52.1% using the recent versions of the reference genome sequence and TE annotations (Tang et al., 2018). Based on the transposition mechanism, there are two classes of TEs: class I transposons, also called retrotransposons, that transpose by copy and paste mechanism, and class II transposons, also called DNA transposons, that transpose by cut and paste mechanism (Deininger et al., 2003; Kazazian, 2004; Stewart et al., 2011). Class II TEs are less abundant in the human genome (3.5%) and are considered DNA fossils (remnants from the ancestral genome) as no family of DNA transposons currently remains active (Pace & Feschotte, 2007). Retrotransposons, therefore, represent the major types of TEs in the human genome due to their replicative transposition and ongoing activity. There are different types of retrotransposons including endogenous retroviruses (ERVs) which are characterized by the presence of long terminal repeats (LTRs), and non-LTR retrotransposons. Non-LTR retrotransposons are further divided into long interspersed elements (LINEs), short interspersed elements (SINEs), SVAs (chimera of SINEs, variable number tandem repeats (VNTRs), and *Alu*-like), and processed pseudogenes (also called retro-genes). Non-LTR retrotransposons are characterized by poly A tail and target site duplication (TSD) with the former unique to this TE type but the latter common to all TEs (Allet, 1979; Grindley, 1978). LINEs have the largest contribution in the human genome at 17.8% followed by SINEs (10.5%), ERVs (9.1%), and SVAs (0.1%). SVAs are very young and active class of TEs despite having only ~5000 copies in the human genome (Tang et al., 2018; Wang et al., 2005). Processed pseudogenes result from retrotransposition of processed protein-coding gene transcripts (mRNAs) with more than 10,000 processed pseudogenes identified in the

 doi:10.20944/preprints202011.0548.v1

human genome (Pei et al., 2012). Formation of processed pseudogenes has been revealed to be ongoing as indicated by the report of at least 48 polymorphic retro-genes in the human genome (Ewing et al., 2013).

The previous notion of TEs being junk or selfish DNA has been revolutionized with the revelation of TEs' role in genome evolution and gene function (Ayarpadikannan & Kim, 2014; Cordaux & Batzer, 2009). TE insertions tolerated during evolution have many effects on structure and function of human genome and shaped the evolution of human lineage (Britten, 2010). Impact of TEs on human genome evolution has been thoroughly discussed in the earlier reviews by Ayarpadikannan and Kim (2014) and Cordaux and Batzer (2009). To recapitulate, TEs are an important factor responsible for rearrangements in the human genome including tandem duplications and insertion- and recombination-based deletions (Bailey et al., 2003; Han, 2005; Sen et al., 2006). Besides large-scale genomic rearrangements, TEs are also involved in local genomic instability and have been found to generate microsatellites in the human genome (Ahmed & Liang, 2012; Kelkar et al., 2007). Another impact of TEs is creation of new genes with functions essential to the host (Elisaphenko et al., 2008; Sha et al., 2000). These molecular domestication events occurred repeatedly during evolution of eukaryotic lineages. For examples, the centromere-associated protein, CENP-B, is derived from a DNA transposon superfamily and is highly conserved across mammals (Smit & Riggs, 1996); *Xist* gene at X-inactivation loci in the eutherian genomes shows dual origin with the exons evolved from *Lnx3* genes and different TEs (Elisaphenko et al., 2008). In the study by Tang et al., (2018) it has been found that more than half of the human-specific TEs (a total of 4,607) are located in protein coding genes, non-coding RNAs (ncRNAs) and transcribed pseudogenes, making up 134 Mbp of reference transcriptome (Tang et al., 2018). Another important function of TEs in the human genome is their involvement in gene expression regulation.

 doi:10.20944/preprints202011.0548.v1

As will be discussed in this review, the two principal methods by which TEs regulate the expression of genes are: function as *cis*-acting regulatory sequences and encoding of regulatory RNAs. Ongoing TE insertions of certain TE subfamilies in the human genome can lead to insertions of TEs in genic regions and alteration in the level of gene expression via different mechanisms including alternative splicing, introduction of premature stop codon, and introduction of polyadenylation and termination signals (Han et al., 2004; Stacey et al., 2016; Vidaud et al., 1993). This can be considered as another way by which TEs can control gene expression level. Our study is however mainly focused on TEs' direct participation in gene regulation via TE-derived *cis*-regulatory regions and TE-derived regulatory RNA sequences in the human genome. In this review, we aim to comprehensively cover the major studies regarding these two aspects of TE-mediated gene regulation in the human genome, and based on these studies' findings to address these questions: What is the extent of TEs' contribution and how versatile is the role of TEs? Does TE-mediated gene regulation tend to be tissue-specific? Does TE-mediated gene regulation lead to evolutionary novelty? How different classes of TEs differ in contributing to gene regulation?

2. *Cis*-Regulatory Activities of TEs

TEs considerably contribute to the *cis*-regulatory regions of the human genome. It has been observed that TEs contribute to almost half of the open chromatin regions (Jacques et al., 2013). Although accessibility does not equate regulatory function, a recent review analyzing the relationship between physical and functional genome concludes that chromatin accessibility plays a wide role in defining active regulatory elements (Klemm et al., 2019). The fact that TEs contribute ~50% of the open chromatin regions demarcates the role of TEs in gene regulation. As established by different studies, TEs either provide alternative promoters and enhancers or increase the activity of existing promoter (Conley et al., 2008; Franchini et al., 2011). The jumping nature

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

along with the presence of intrinsic regulatory sequences in TEs for their own expression as well as TEs' susceptibility to recruit silencing factors for their own suppression, make them a crucial player in controlling gene expression pattern. This section of the review will cover TEs' *cis*-regulatory activities, including TEs' involvement in important gene regulatory elements, genes that have been found to be controlled by TEs' regulatory activities, spatial gene regulation by TE-derived *cis*-regulating elements, conservation of the TEs-derived *cis*-acting elements across species, and polymorphic TEs leading to population-specific gene expression patterns.

2.1 Contribution of TEs in different regulatory elements in the genome

2.1.1 Regulatory elements in the genome

Cis-regulatory regions (including promoters, enhancers, silencers, and insulators) are non-coding DNA sequences that regulate gene expression by providing binding sites for *trans*-acting factors. Promoters are orientation-dependent regulatory elements with respect to the genes and provide a docking site for basic transcriptional machineries. Other regions that control transcription in the eukaryotic genome include enhancers, silencers and insulators. Unlike promoters, enhancers and silencers are orientation- and position-independent with respect to genes. Enhancers typically consist of clusters of transcription factor binding sites (TFBSs) that work cooperatively to up-regulate gene expression. Silencers in contrast down-regulate gene expression by recruiting factors that promote close chromatin structures. Insulators are another type of regulatory elements that protect genes from the regulatory influence of the surrounding genes. All of these regulatory regions in the genome play a crucial role in gene regulation by interacting with a wide range of *trans*-acting factors.

 doi:10.20944/preprints202011.0548.v1

Databases of gene regulatory regions: To provide a comprehensive map of gene regulatory regions in the human genome, different approaches have been used, including identification of open chromatin regions, localization of binding sites of transcription factors (TFs) and other gene regulatory proteins and mapping of the chromatin states by identifying the sites of DNA methylation and active and repressive histone marks (Bernstein et al., 2010; Gao & Qian, 2019). In order to acquire these datasets, a wide range of high-throughput functional genomics techniques have been utilized. For identification of open chromatin regions in the genome, the commonly employed DNA accessibility assays include DNase-seq and FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements)-seq (Giresi et al., 2007; Song & Crawford, 2010). For identification of TFBSs and binding sites of epigenetically modified histones, ChIP (Chromatin Immunoprecipitation)-seq technique is used (Robertson et al., 2007). For mapping of DNA methylation sites in the genome, WGBS (Whole Genome Bisulfite Sequencing) and RRBS (Reduced Representation Bisulfite Sequencing – that only targets promoters/CpG islands) are the commonly employed assays (Kernaleguen et al., 2018). There are different databases that provide gene regulation datasets by either reporting data of these experiments separately or by integrating the data of different assays to define promoter and enhancer elements in the genome. Two important databases providing the massive data of the functional genomics experiments mentioned above are ENCODE (encyclopedia of DNA elements) project database ("The ENCODE (ENCyclopedia Of DNA Elements) Project", 2004) and REMC (Roadmap Epigenomics Mapping Consortium) project database (Bernstein et al., 2010). The data encompass a wide range of tissues and cell lines. Some of the small-scale projects are GGR (Genomics of Gene Regulation) that includes data mainly for the A549 cell line and few primary cells, and the blueprint epigenome project database (Martens & Stunnenberg, 2013) providing data for distinct types of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

haematopoietic cells. Based on these primary datasets, there are some secondary databases to provide meaningful interpretation of the primary data in various ways. For example, an enhancer database, EnhancerAtlas (Gao & Qian, 2019), provides enhancer annotations across nine different species by combining output of multiple high-throughput experiments. It integrates the ChIP-seq datasets of histone modifications, TFs, and other regulatory proteins that specifically bind to enhancers; different open-chromatin datasets (DNase-seq, FAIRE-seq, and MNase-seq), as well as the findings of some reporter assays to demarcate enhancer regions in the genome. Another enhancer database is SEdb (Jiang et al., 2019), which is a comprehensive database of super-enhancers (large cluster of transcriptionally active enhancers) in the human genome. Table 1 summarizes primary and secondary gene regulation databases.

Table 1: Comprehensive list of major primary and secondary gene regulation databases

Primary databases			
Database	**Brief description**	**Specie**	**Reference**
ENCODE (Encyclopedia of DNA Elements)	Provides following functional genomics data for the diverse range of tissues and cell lines: DNase-seq data, FAIRE-seq data, Histone ChIP-seq data, TF ChIP-seq data	Human	("The ENCODE (ENCyclopedia Of DNA Elements) Project", 2004)
REMC (Roadmap Epigenomics Mapping Consortium)	Provides following functional genomics data for the diverse range of tissues and cell lines: DNase-seq data. Histone ChIP-seq data, WGBS data, RRBS data	Human	(Bernstein et al., 2010)
GGR (Genomics of gene regulation)	The database is limited to only A549 cell line and few primary cells. Provides following functional genomics data: DNase-seq data, Histone ChIP-seq data, TF ChIP-seq data	Human, Mouse	https://www.genome.gov/Funded-Programs-Projects/Genomics-of-Gene-Regulation
Blueprint epigenome project	Provides reference epigenomes of distinct types of haematopoietic cells. Includes following functional genomics data: DNase-seq data, Histone ChIP-seq data, WGBS data	Human	(Martens & Stunnenberg, 2013)
Secondary databases			
Database	**Brief description**	**Specie**	**Reference**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

OCHROdb (**O**pen **Chro**matin Database)	Integrates DNase seq data from ENCODE, Roadmap Epigenomics, Genomics of Gene Regulation and Blueprint Epigenome to provide comparison of open chromatin regions across multiple samples	Human	(Shooshtari et al., 2018)
ChIPSummitDB	Determines cistrome of TFs by analyzing TF ChIP-seq data from primary databases.	Human	(Czipa et al., 2020)
SEdb (Super-enhancer database)	Maps super-enhancer regions in the genome by analyzing ChIP-seq data of H3K27ac. The current version documents a total of 331 601 super-enhancers from 542 samples.	Human	(Y. Jiang et al., 2019)
EnhancerAtlas	Identifies enhancer region by integrating datasets of 12 high-throughput methods. In contrast to other enhancer databases (SEdb, HACER, RAEdb, HEDD, DiseaseEnhancer, TiED, GeneHancer, SEA, DENdb and dbSUPER), it combines versatile and most comprehensive set of annotations.	The latest version has data for 9 species including human	(Gao & Qian, 2019)
Genome Segmentations from ENCODE data	Identifies functional regulatory elements in the genome by integrating ChIP-seq data for 8 chromatin marks, RNA Polymerase II, the CTCF transcription factor. It involves application of two unsupervised machine learning techniques (ChromHMM and Segway) to assign genomic states to disjoint segments in the genome.	Human	(Ernst & Kellis, 2012; Hoffman et al., 2013)
Cistrome DB (Cistrome Data Browser)	Combines raw ChIP-seq and chromatin accessibility data from ENCODE, Roadmap and few other resources and process it through same pipeline and quality control metrics to achieve consistency and provides a dataset with standardized curation, quality control and analysis procedures.	Human, mouse	(Mei et al., 2017)

2.1.2 Intrinsic regulatory properties of TEs

Many studies have revealed that TEs contribute to all regulatory regions described above (Brini et al., 1993; Franchini et al., 2011; Hambor et al., 1993; Samuelson et al., 1990). Intrinsic regulatory properties of TE sequences make them a suitable candidate for regulating gene expression. Like other genes, TEs may harbor the primary types of regulatory sequences for their own expression: promoters, enhancers/insulators, splice sites, and terminators. Internal regulatory sequences of the retroelements can carried into the progeny copies (Swergold, 1990; van Regenmortel & Mahy, 2010). LTRs and LINEs carry POL II promoters while SINEs carry promoters for either POL III or POL II (Roy et al., 2000; Swergold, 1990). SVAs contain core enhancer element (Khoury & Gruss, 1983) within the SINE-R sequence (Ono et al., 1987). According to one of the models proposed for SVA transcription, the internal enhancer element of SVAs acts cooperatively with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

the external promoters to promote SVA transcription (Hancks & Kazazian, 2010). While processed pseudogenes derived from mRNA lack promoter sequences, the transcriptional regulatory signals of 5' untranslated region (UTR) might impart intrinsic regulatory functions to these retrogenes (Dikstein, 2012). Nevertheless, not all TEs are equally co-opted as gene regulators, and their abundance and replication mechanism determine the fate of their regulatory functions after transposition as new copies or at new locations. For example, LTR retroelements retain their regulatory sequences in the genome after the insertion, even though their coding sequences may be deleted via frequently occurred homology-based recombination deletion between their LTRs. In contrast, LINEs frequently undergo 5' truncation during retrotransposition, resulting in the deletion of its promoter regions (Chuong et al., 2017).

2.1.3 TEs contribute to regulatory elements in the genome

TEs exaptation to regulatory elements in the human genome has been well documented. For examples, Franchini et al. (2011) discovered that an LTR retrotransposon (belonging to THE1B in the MaLR subfamily) exaptation causes evolution of an enhancer element, which leads to neuronal specific expression of *POMC* gene in mammals. LTR retroelements of the same subfamily have also been found to be involved in abnormal expression of *CSF1R* gene in Hodgkin lymphoma. In this case, transcription of *CSF1R* in transformed human cells was found to be initiated at an anomaly activated LTR retroelement (Lamprecht et al., 2010). Another study showed that the insertion of an ERV repeat in the upstream region of *AMY1* gene leads to the activation of cryptic promoters and tissue-specific expression of the gene (Samuelson et al., 1990). Two reports established the role of *Alu* elements in evolution of T cell promoters and enhancers: an *Alu*Sp in the promoter of *FCER1G* gene induces T cell expression; an *Alu*Y in the intron of human *CD8* gene acts as a T cell enhancer. Both these *Alu* sequences harbor the binding motifs of Lyf-1 TF,

 doi:10.20944/preprints202011.0548.v1

which drives T cell-specific expression (Brini et al., 1993; Hambor et al., 1993). Transcription of the *Alu*Sq from its POL III promoter prevents the human epsilon globin gene from regulation by the activities of the other upstream promoters, showing *Alu* as an insulator (Wu et al., 1990). Another study identified a TFBS for retinoic acid response element (RARE) in an *Alu*S inserted in the promoter of human *KRT18* gene, leading to differential expression in tissues (Neznanov et al., 1993). A study by Kim et al. (2011) identified alternative promoters derived from L1 and SVA elements in *CHRM3* and *WDR66* genes, respectively (Kim & Hahn, 2011).

Recently, the contribution of TEs in the promoters of genes expressed by POL II was determined using ENCODE and RepeatMasker annotations for TFBSs and TEs, respectively, by analyzing promoters as the 1500 bp regions upstream of the transcription start sites (TSSs) (Kellner & Makałowski, 2019). Out of the 35,007 promoters, 75% were found to have TE-derived sequences with some promoters found to have as many as ten TEs. The percentage is almost similar to the result of a previous study (83%) that took into account the 2000 bp sequences upstream of TSSs (Thornburg et al., 2006). The difference might be due to the variation in the length of upstream sequences analyzed, as it has been reported that TE density increases with the distance from TSS (Kellner & Makałowski, 2019). In a recent work by Zeng et al., (2018), TE enrichment was determined in different regulatory regions by measuring 'P(TE|RE)', the probability of nucleotide in the regulatory element being from the TE. Interestingly, P(TE|RE) was found to be higher in repressors than promoters, reaching 0.2 and 0.5 for promoters and repressors, respectively (Zeng et al., 2018). The role of TEs as gene repressors has also been supported in other studies that showed that TEs can repress nearby genes by spreading local heterochromatin (Brattås et al., 2017; Liu et al., 2018). The study by Brattås et al, investigating the ERV expression pattern in human brain revealed that TRIM28, a corepressor protein, binds on the docking site on ERV and

 doi:10.20944/preprints202011.0548.v1

consequently regulates the nearby genes (Brattås et al., 2017). L1-mediated transcriptional repression of neighboring genes has also been observed in human cell lines (Liu et al., 2018).

In summary, TEs' co-evolution with the regulatory elements in the genome has thus been well-established. Studies have revealed TE sequences embedded in regulatory elements, as well as the regulatory role of these TEs. Besides their contribution in canonical promoters, TEs have also been found to create alternative promoters for certain genes. From the studies mentioned in this section, it can be concluded that TEs are the reservoir of diverse regulatory functions and play an important role in evolution of different types of regulatory elements.

2.1.4 Contribution of TEs to TFBSs

Studies have documented the binding of TFs to TEs and showed TEs have TF-binding sequence motifs (Kellner & Makałowski, 2019; Sundaram et al., 2014; Sundaram & Wysocka, 2020). TE sequences widespread in the human genome could provide binding sites for many classes of TFs (Kellner & Makałowski, 2019). In the study by Sundaram et al., (2014), TF binding regions (TF ChIP-seq binding peaks) of 26 TFs were analyzed in two human cell lines (K562 and GM12878), and it was observed that 20% of the TF binding peaks belonging to wide range of TFs were found to be derived from TEs (Sundaram et al., 2014). TEs contribute to TFBSs by providing ready-to-use TFBSs immediately after insertion and by generating novel TFBSs via post-insertion random mutations. Presence of TF-binding motifs in TEs prior to their insertion has been indicated in the work conducted by Ito et al. (2017). The study determined TFBSs in the LTR retroelement (HERV-TFBSs) and later determined TF-binding motifs that were found in a substantial fraction of HERV-TFBSs at the same consensus position (named 'HERV/LTR-shared regulatory element – HSRE' by the author). HSREs were found in 2% of all the TFBSs in the genome (Ito et al.,

 doi:10.20944/preprints202011.0548.v1

2017). In addition to the use of existing TFBSs, creation of TFBSs in TEs after their insertion has also been reported. For example, methylated CpGs of human *Alu* sequences can undergo deamination (C->T mutation) to create binding site for c-Myc TF (Zemojtel et al., 2011). Another study revealed that a single C to T substitution in the *Alu* sequence leads to functional binding site for Lyf-1 TF (Hambor et al., 1993). Deamination of CpG in *Alu* sequences has also been found to originate binding sites for RAR (Rayan et al., 2016). Likewise, deamination of methylated CpG sequences to TpG in human LTRs has been shown to create binding sites for p53 (Zemojtel et al., 2011). The role of mutations in TEs in providing new regulatory sequences is supported by genome-wide studies analyzing TE-derived TSSs in the human transcript libraries, which showed that old L2 elements are more likely to contribute to promoters than new L1s (Faulkner et al., 2009).

Occurrence of TFBSs across TEs in the human genome is not random. Binding sites of a TF are enriched in copies of specific TE families. A total of 710 such TF-TE relationships have been identified (Sundaram et al., 2014). Non-random association of TEs with TFBSs is also indicated by TEs providing combinatorial interaction of TFs. TEs provide clusters of binding sites for TFs that work cooperatively in gene regulation. For example, MIR family of SINEs that have affinity for estrogen receptor α (ERα) also provide binding sites for ERα co-factors (Testori et al., 2012). The non-random association of TEs with TFBSs signifies the role of TEs in shaping gene regulation networks.

TEs are considered as a source of a large number of TFBSs in the human genome. It has been observed that TFs with a greater number of TF ChIP-seq peaks not only have a greater number of TE-derived peaks as expected, but also have a greater fraction of TE-derived peaks indicating TEs being responsible for generation of TFBSs (Sundaram et al., 2014). In another study analyzing the

 doi:10.20944/preprints202011.0548.v1

role of genome expansion in the evolution of gene regulation (Marnetto et al., 2018), it has been indicated that TFs increase their targets in the genome through genome expansion mainly by repeat elements. The study determined the age of human genomic regions and their TFBS distribution by applying parsimony model to genome-wide alignment of 100 vertebrates. It was found that binding sites of a TF were enriched in genomic regions of a given age, which suggests that new genomic sequences provide new targets for existing TFs (Marnetto et al., 2018). In concordance with the role of TEs in expanding TFBSs, TE-derived TFBSs are considered as the marker of gene regulation evolution. In the study conducted by Nikitin et al., (2019), evolution of transcriptional regulation was determined for different genes and pathways using retroelement-derived TFBS as a metric. Genes enriched for TE-derived TFBSs and the associated pathways were considered to have high evolutionary rates.

Functional significance of TE-derived TFBSs in the human genome has been highlighted in several folds. First, functionally important positions of TE-derived TFBSs that interact with TFs are more conserved than other adjacent positions, indicating the sign of functional constraints on these TFBSs (Polavarapu et al., 2008). Second, TEs that are de-repressed in cancers have been found to harbor binding sites for oncogenic TFs including C/EBPβ, E2F1, and MYC (Jiang & Upton, 2019). In the study conducted by Kellner and Makałowski (2019), 6.8% of TFBSs present in the promoter elements were found to be derived from TEs. Presence of TE-derived TFBSs in the promoter regions indicates their regulatory function. Moreover, TE sequences not associated with genes but harboring TF binding motifs could participate in gene regulation by acting as competitors of the genes' regulatory sequences in binding to TFs.

2.1.5 Differential contribution of TEs by type in regulatory regions

 doi:10.20944/preprints202011.0548.v1

The contribution of TEs to the regulatory elements in the human genome varies among TE types. The study by Zeng et al. (2018) determined the proportion of nucleotides belonging to different types of TEs in regulatory regions. It revealed that *Alu* elements contribute most to all types of regulatory regions, while L1s were found to be least likely in the regulatory regions. The authors of the study reasoned that the large size of L1s and even truncated L1 copies might disrupt the genic regions of the genome, and therefore L1 insertions in the regulatory elements have not been evolutionary favored. Furthermore, as L1s on average are older than *Alu* elements, a more significant contribution of *Alu* elements than L1s in different types of regulatory elements was considered as indicative of the idea that clade-specific and species-specific TEs are more likely to contribute in gene regulation. This finding is also supported by the study of Nikitin et al. (2018), which revealed that SINE-derived TFBSs are more in number than LINE-derived TFBSs in gene neighboring regions (5 Kb surrounding TSS), while it is the way around for regions outside the gene neighborhood. Another support has been provided by the recent study by Kellner and Makalowski (2019), which indicated that SINEs are more frequent in promoters (1.5 Kb upstream of TSS) than non-promoter regions, while it is the opposite for LINEs. Hence, multiple studies have shown in different ways that SINEs contribute more to regulatory regions than LINEs.

Although the presence of *Alu* elements in regulatory elements signifies the role of lineage-specific TEs in gene regulation, it has been found that ancient repeat elements including L2 and MIRs show a higher nucleotide proportion in enhancers despite having lower sequence contribution to the genome (Zeng et al., 2018). In another study, analysis of TE-derived TFBSs showed that ancient TE families like MIRs and L2s are more enriched for TE-derived TFBSs (have more TFBSs than expected based on their genome frequencies) than younger families like *Alu* elements and L1s (Polavarapu et al., 2008). As suggested by the authors, the presence of ancient TEs in these TFBSs

 doi:10.20944/preprints202011.0548.v1

highlights the functional conservation of TE-originated regulatory sites (Polavarapu et al., 2008). Based on these findings, it can be said that although the exaptation of younger TEs to regulatory elements evolves gene regulation, certain classes of regulatory elements are enriched for older TE families indicating functional conservation of TE-originated regulatory sites.

Besides SINEs and LINEs, LTRs are also considered as an important TE class playing a role in gene regulation as they retain their regulatory sequences after their integrations, and they are the most dominant TE class in open chromatin regions of the human genome (Jacques et al., 2013). Moreover, ERVs/LTRs are the most diverse class of human TEs, providing various regulatory elements and TFBSs (Ito et al., 2017). The study by Thornburg et al., also showed that unlike LINEs, SINEs and DNA elements, LTRs are enriched for binding sites of the majority of TF classes (Thornburg et al., 2006). Investigating the regulatory properties of different classes of LTRs has therefore remained an important area in TE-mediated gene regulation. However, studies analyzing the number of TE-derived TFBSs for different types of TEs in upstream gene regions, have not found the major contribution of LTRs, which implies that LTRs might be involved in regulating distant genes. These studies analyzing upstream gene regions for different TE types have revealed that SINEs are the major contributors followed by LINEs and then LTRs (Kellner & Makałowski, 2019; Nikitin et al., 2018).

In summary, we reviewed in this section TEs' contribution to the major regulatory elements in the genome, highlighting some important functional aspects of TE-mediated gene regulation like activation of cryptic promoters by TEs and combinatorial interactions of TFs contributed by TEs. The role of TEs has been observed in promoters, enhancers, and silencers in the human genome. This diversity of TE-mediated gene regulation can be linked to a wide variety of TFBSs provided by TEs and different types of intrinsic regulatory properties present in TEs for their own regulation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

Nevertheless, studies experimentally verifying the functional role of TEs in regulatory elements are still limited, and future work in this direction can employ reporter gene expression under the control of promoters with and without the TE-derived sequences to elucidate TEs' specific roles in gene regulation.

2.2 Genes regulated by TE-derived *cis*-regulatory sequences

Many genes in the human genome have their expression known to be controlled by TE-derived regulatory sequences. Some studies focusing on specific genes have identified TE-derived regulatory elements by using a reporter gene expression approach or by identifying alternative transcripts initiated at TE sequences. A few of these studies were already highlighted in the previous sections, and as examples, *POMC*, *CSF1R*, *FCER1G*, and *CD8* genes are regulated by TE-derived regulatory elements (Brini et al., 1993; Franchini et al., 2011; Hambor et al., 1993; Lamprecht et al., 2010; Samuelson et al., 1990).

Genome-wide analysis has also been conducted by different research groups to identify TEs in the gene upstream regulatory elements. The study by Kellner et al. (2019) showed that 75% of the 35,007 genes transcribed by POL II have TE-derived sequences in their promoter regions, which represents enrichment over the genome average. This coincides with the TEs' preferential insertion in the upstream gene regions (Sultana et al., 2017). The same study further identified that for two protein-coding genes, *PCBD1* and *PPP1R3A*, almost the entire promoters are derived from TE sequences (Kellner & Makałowski, 2019). The study by Nikitin et al. (2018) showed that among the protein-coding genes, *USP176L26*, *USP17L13*, and *USP17L12* genes (encoding ubiquitin associated peptidase) most strongly associate with TE-derived TFBSs.

 doi:10.20944/preprints202011.0548.v1

TEs can also regulate the far away genes by acting as enhancer elements. Raviram et al. (2018) analyzed 3D genomic interactions to determine the genes regulated by ERVs. They used Chromosome Conformation Capture (3C) methodologies to determine the transposons' contribution to chromatin folding and long-range intra-chromosomal interaction and provided a strategy to identify TE-regulated genes, specifically genes interacting with TE-derived enhancers. It was found that the *IF16* gene is up-regulated by a retroelement MER41B. The gene promoter was found to be interacting with this LTR located ~20 Kb downstream of the gene. Similarly, the technique captured the interaction between *IFITM* (*IFITM1* and *IFITM3*) genes and MER41A retrotransposons located downstream of the genes. Expression of the *MYPN* gene (specifically expressed in heart and skeletal muscle) was also found to be regulated by distant TE enhancers (Raviram et al., 2018). All these examples signify the importance of unveiling long-range genomic interaction of TEs in identifying TE-regulated genes.

In summary, expression of a certain number of genes has been experimentally validated to be controlled by TEs, followed by recent genome-wide data analytical studies that have revealed TE sequences in many genes regulatory regions underscoring the need to further investigate the topic. Genes with TE-derived regulatory sites have a wide range of functions. Among many other products, these genes encode for neuropeptides (POMC), muscle protein (MYPN), immune receptors (FCER1G and CD8), metabolic enzymes (AMY1), and signaling receptors (CSF1R). The functional diversity of the genes being regulated by TEs indicates TEs' diverse impact on the host phenotype. Further, as to be discussed in detail later, some studies also showed that genes crucial for speciation novelty have TEs in their regulatory regions, highlighting the importance of TEs in evolution and functional diversity.

2.3 Tissue-specific gene regulation by TEs

 doi:10.20944/preprints202011.0548.v1

The epigenetic status of TEs varies across human tissues (Pehrsson et al., 2019), leading to the varying profile of TE regulatory activities in different tissue types. Tissue-specificity is considered as one of the ways, in which TEs contribute to evolutionary novelty in gene regulation. Studies focusing on specific genes have revealed TEs' exaptation to tissue-specific regulatory sequences. For examples, as mentioned before, an LTR retroelement provides neuronal enhancer of *POMC* gene, and *Alu* sequences were found to provide T cell promoter and enhancers for *FCER1* gene and *CD8* gene, respectively (Brini et al., 1993; Franchini et al., 2011; Hambor et al., 1993).

Genes with LTR retroelement in the upstream regions have been found to exhibit tissue-specific expression compared to LTR-unassociated genes (Pavlicev et al., 2015). In this systematic study, gene expression data of 18 different tissue types were analyzed from Illumina Human Body Map 2.0 (HBM2.0). The study determined co-expression (a metric of distinctive gene expression pattern in a tissue compared to other tissues) of LTR-associated and LTR-unassociated genes and found 62 LTR elements linked to tissue-specific gene expression (Pavlicev et al., 2015). Trizzino et al. (2018) used the data of the 'Roadmap Epigenomics Project' and 'Genotype tissue expression project' to determine TEs in active and repressed chromatin of different tissues and the consequences on the gene expression. Interestingly, genes having the same expression in different tissues (i.e., lack of tissue-specific expression) rarely had TE insertions in their regulatory regions. It was found that TEs' (particularly LTRs) involvement in the active chromatin regions varies across tissues. For instance, HERV15 is significantly enriched in active chromatin of liver tissue, while X7C (LINE) and Charlie15a (DNA transposon) are enriched in the active chromatin of breast tissue. Further, the tissue-specific TE involvement in active chromatin was linked to tissue-specific gene expression. It was revealed that TEs in the active chromatin regions of a tissue have binding sites for that tissue's key TFs. For example, HERV15 is more enriched in the active chromatin

 doi:10.20944/preprints202011.0548.v1

regions of the liver, and it has binding sites for EOMES, which is the key TF in hepatic immune response. The tissue-specific involvement of TEs in active chromatin regions was also found to be associated with altered gene expression levels in that tissue (Trizzino et al., 2018). The study by Kellner and Makalowski, (2019) examined the ENCODE data of TFBSs in six different tissues (blood, breasts, kidney, liver, lung, and stem-cells) in a pair-wise fashion and found that only a small fraction of TE-derived TFBSs active in one tissue was used in another tissue. For example, only 3% of TE-derived TFBS active in blood tissue was also used in breast tissue. For almost all the tissue pairs, this percentage was significantly smaller for TE-derived TFBSs than for all TFBSs, indicating the role of TEs in tissue specificity of gene expression. To give an example, 9% of all the TFBSs active in blood tissue was also active in breast tissue but just 3% of the TE-derived TFBSs active in blood tissue were also used in breast tissue (Kellner & Makałowski, 2019). Moreover, in a very recent study analyzing ENCODE data for human GM12878 and K562 cell lines, it was shown that variability in the TE-derived CTCF sites across different cell types leads to chromatin looping variation and alternative promoter-enhancer interactions associated with the difference in gene expression across cell types (Diehl et al., 2020).

As highlighted by the studies mentioned above, tissue-specificity of TE-mediated gene regulation has been corroborated using different approaches. Sequence analysis of some of the reported tissue-specific enhancers revealed that they harbor TEs. Furthermore, tissue gene expression data showed that genes with TEs in upstream regions have a distinct expression in tissues, which has been subsequently supported by studies that determined tissue-specific TE-derived TFBSs and active chromatin regions. In summary, many TEs providing *cis*-regulatory sequences tend to function in a tissue-specific fashion and play an essential role in the differential gene expression across tissues.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

2.4 Lineage-specific gene regulation by TEs

TEs have been observed in the lineage- and species-specific regulatory regions implying the role of TEs in evolving gene regulation. The work by Rosario and co-workers determined that 56% of the anthropoid-specific regulatory elements have a TE origin (Rayan et al., 2016). In the study by Trizzino et al. (2017), human liver promoter and enhancer sequences were compared across six primate species and it was found that the majority of the non-conserved elements are enriched in TEs including LTRs and SVAs (Trizzino et al., 2017). The emergence of TE-derived lineage-specific regulatory sites is either due to newly evolved species-specific TEs or might be due to mutations in the ancestral TEs (Faulkner et al., 2009; Kunarso et al., 2010; Lynch et al., 2011) (Figure 1). The creation of gene regulatory sites by mutations in the ancestral TE sequences is supported by the finding that most of the TEs in the regulatory regions have a high sequence divergence (>8% diverged) (Nikitin et al., 2018). This has also been considered as the reason behind the higher contribution of ancestral TE families (L2 and MIR) than that of L1 and *Alu* in some regulatory regions, as mentioned before in section 2.1.4 discussing the generation of new TFBSs in the genome by mutations in TE sequences. Moreover, lineage-specific TEs are also the source of lineage-specific TE-derived regulatory sites. Different vertebrate lineages contain quantitatively and qualitatively different populations of TEs, essentially due to different evolution of ancestral families of TEs, the lineage-specific introduction of TEs by infection, and lineage-specific emergence of new TEs subfamilies, as well as ongoing transposition from existing active TEs. Lineage-specific TEs have been revealed to participate in lineage-specific gene regulatory regions. A study showed that only 5% of TFBSs for *Oct4* and *Nanog* (key regulators of embryonic stem cells) are conserved between human and mouse embryonic stem cells, and the majority of the non-conserved sites reside within species-specific LTRs (Kunarso et al., 2010). This links the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

emergence of species-specific TEs to the evolution of gene regulatory networks involved in pluripotency and cell fate determination. Another study indicates the role of transposons in gene regulatory networks crucial for speciation novelty (e.g., pregnancy in eutherian mammals). It was found that 13% of the genes showing endometrial expression in placental mammals had eutherian-specific TEs in the upstream region (Lynch et al., 2011). Moreover, it has been found that in the human genome, 30% of the TFBSs of the tumor suppressor protein, p53, reside in the primate-specific ERV regions (Wang et al., 2007). The findings of these studies show that the emergence of species/lineage-specific TEs contributes to the evolution of gene regulatory network pertinent to significant biological functions, including pluripotency of ESCs, lineage-specific traits like pregnancy in placental mammals and tumor suppression.

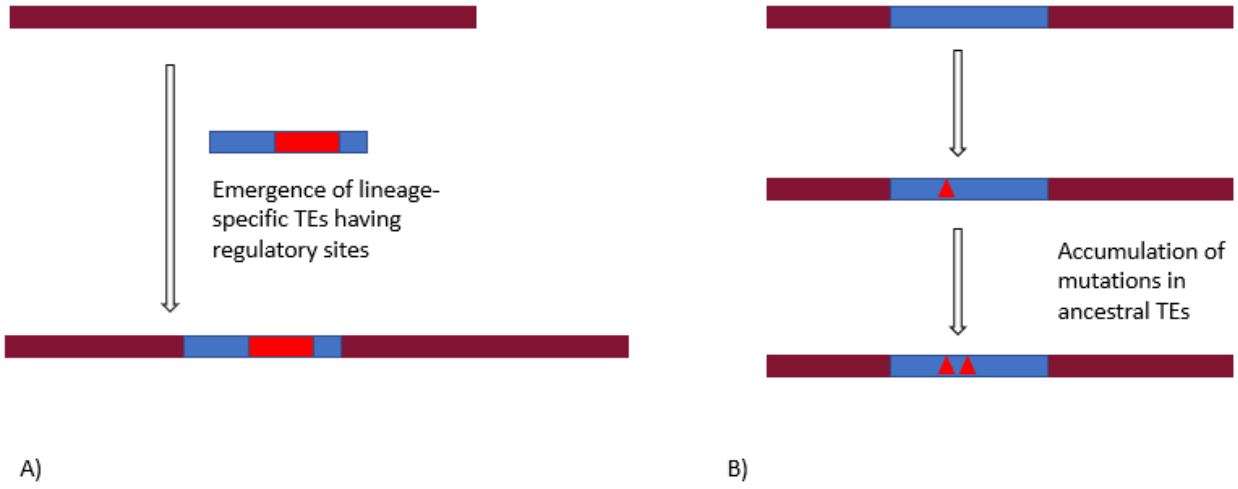

Figure 1: Two different pathways of generating lineage specific TE-derived regulatory sites. Lineage specific TE derived regulatory sites arise due to emergence of lineage specific TEs in the genome (A), or might be due to accumulation of mutations in ancestral TEs (B)

The higher contribution of ancestral TE subfamilies (L2 and MIR) than L1s and *Alu* elements in some regulatory regions might seem contradictory to the lineage specificity of TE-mediated gene

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

regulation. However, as mentioned before, sequence divergence of ancestral TEs evolves regulatory regions in species. Nevertheless, TEs indeed have also been identified in the conserved mammalian-wide regulatory elements. For example, a neuronal-specific TE-derived enhancer of the *POMC* gene exapted before the origin of prototherians (~166 Mya) (Franchini et al., 2011). Concludingly, besides providing conserved regulatory functions, TE-derived regulatory sites also tend to be specie/lineage-specific and contribute to speciation novelty and diversity. Future comprehensive analysis encompassing all categories of regulatory elements across a wide range of species should provide more insight.

2.5 Population-specific gene regulation by polymorphic TEs

The majority of the TEs in the human genome are fixed and derived from ancient transposition events. Previous studies exploring the regulatory effects of TEs mostly focused on the ones fixed in the human population. Nevertheless, mobile element insertion (MEI) polymorphisms have been found to be the most frequent structural variants in the human genome. The three families of retrotransposons primarily responsible for generating human TE polymorphisms are *Alu* elements, L1s, and SVAs (Auton et al., 2015; Batzer & Deininger, 1991; Brouha et al., 2003; Wang et al., 2005). It is estimated that two haploid human genomes differ by about 1000 TEs insertions, and thus the reference human genome does not represent a complete archive of human TEs (Bourque et al., 2018). More than 16,000 polymorphic TE loci were identified in the recent phase 3 variant release of the 1000 Genome Project (Auton et al., 2015). Furthermore, a recent analysis of deeply sequenced whole genome data of 152 populations from ‘The Simon Genome Diversity Project’ discovered more than 5000 additional MEIs not reported by the 1K genome project (Watkins et al., 2020). Based on TEs’ intrinsic regulatory activity, it is very likely that polymorphic TEs are involved in differential gene expression among human populations by offering new regulatory

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

sites to their nearby genes. Interestingly, studies have shown that many polymorphic TE loci in humans correspond to *cis*- and *trans*-eQTLs (Spirito et al., 2019; Wang et al., 2017). A study by Wang et al. (2017), investigated the association between polymorphic TE loci and gene expression level. Genotype calls for polymorphic TEs were taken from the phase 3 variant release of the 1000 Genomes Project, and corresponding RNA-seq data for the same 1000 Genome Project samples were retrieved from the GUEVADIS RNA-seq project (Lappalainen et al., 2013). It was found that polymorphic TE loci were associated with differences in expression between European and African population groups. A single polymorphic TE locus was indirectly associated with the expression of numerous genes via the regulation of the B cell-specific TF (Wang et al., 2017). In a recent extension of this work (Spirito et al., 2019), rare and less common TE structural variant (TEV) polymorphisms (MAF < 5%) were also included and in total 323 significant TEV-*cis*-eQTL associations were found.

So far, there have been not many studies relating human polymorphic TEs with gene expression differences among populations. The work is limited to only five populations of the 1000 Genome Project data, as only for these populations, the corresponding RNA-seq data is available. Moreover, only lymphoblastoid cell gene expression level has been analyzed. There is a need for more detailed studies encompassing different tissue types and better population coverage to investigate further the correlation between polymorphic TEs and population or even individual level gene expression differences.

3. TEs contribute to non-coding regulatory RNAs

Advancement in RNA-seq technologies has dramatically increased the discovery of new RNAs, the ncRNAs in particular (Derrien et al., 2012; Habegger et al., 2011; Wang et al., 2009). The wealth of ncRNAs is indicated by the fact that about 75 – 85% of the human genome gets

 doi:10.20944/preprints202011.0548.v1

transcribed despite only ~1.2% of the genome encoding proteins (Djebali et al., 2012). ncRNAs include housekeeping RNAs (rRNA, tRNA, snRNA, and snoRNA) and regulatory RNAs (small non-coding RNA (sncRNA) and long non-coding RNA (lncRNA)). Examples of sncRNAs are miRNAs and piRNA. miRNA plays an important role in gene regulation by interacting with the complementary sequence on the 3' UTR of target mRNA, which leads to the cleavage or translation repression of the target mRNA. lncRNAs are further classified based on the genomic region they get transcribed: 1. LincRNAs transcribed from the intergenic regions; 2. Intronic lncRNAs transcribed from introns; 3. lncRNAs that are antisense transcripts of coding regions but do not encode proteins; 4. Circular lncRNAs that have scrambled exon sequences (due to exon shuffling) but do not encode proteins. A plethora of lnc/sncRNA genes have been identified. A total of 15,941 lncRNA and 9882 sncRNA genes have been documented in Gencode v24 (Jalali et al., 2016).

snc/lncRNAs participate in a wide range of regulatory functions by either inducing degradation of mRNA transcripts or regulating the transcription. There is a close association of TEs with regulatory RNAs, as a significant number of these ncRNAs have originated from TEs. This section of the review will highlight TEs' contribution to the regulatory RNAs, mainly focusing on the role of TEs in the origin, functionality, and diversification of regulatory RNAs.

3.1 Contribution of TEs to the makeup of regulatory RNAs

miRNAs are transcribed from genes as primary miRNAs (pri-miRNAs), which are further processed to precursor miRNAs (pre-miRNAs). These initial forms of miRNAs have a stem-loop structure which is later cleaved to form mature miRNA. Mature miRNA is further loaded on argonaute protein to perform gene silencing function (Azlan et al., 2016; Peters & Meister, 2007). Studies have reported the involvement of TEs in the origin of human miRNAs, particularly the

 doi:10.20944/preprints202011.0548.v1

stem-loop structure of different miRNAs families. Supported by the TE-origin of many miRNAs, it has been hypothesized that two similar TEs flanking a genomic locus lead to the formation of miRNA stem-loop structure (Hadjiargyrou & Delihas, 2013). Another study reported an observation of a high sequence identity between the miRNAs of the hsa-mir-548 family and the miniature inverted repeat transposable elements (MITEs). MITE forms a stem-loop structure, which can be recognized by RNAi enzymes and processed into mature miRNA (Piriyapongsa & Jordan, 2007). In the study by Yuan and colleagues (2010), it was shown that the MER53 element, a TE characterized by the presence of terminal inverted repeats (TIRs) and TA target site duplications that can form palindromic structures, gave rise to all members of the miR-1302 gene family (Yuan et al., 2010). In another study, analysis of human palindromic MER sequences using miPred (a tool that distinguishes real miRNA precursor from other hairpin sequences) identified three miRNAs derived from a MER96 located on chromosome 3 and MER91C paralogs located on chromosome 8 and chromosome 17 (Ahn et al., 2013).

TEs have been found to have overlap with pre-miRNA sequences as well as in mature miRNAs. Small RNA sequencing coupled to argonaute2 RNA immunoprecipitation (that captures mature miRNAs) has determined TE-derived miRNA sequences. In a recent study by Petri et al. (2019), TE-derived miRNAs in human brain tissues were identified by conducting Argonaute2 RNA immunoprecipitation followed by small RNA sequencing (AGO2 RIP-seq). The study determined a total of 19 miRNAs that were derived from L2. It was speculated in the study that these L2-miRNAs could target many protein-coding genes carrying L2 sequences in their 3’ UTRs (Petri et al., 2019). Many bioinformatics studies are highlighting the overlap of TEs with miRNA genes. miRBase is a publicly available online repository for miRNA sequences and annotations, allowing researchers to examine the contribution of TEs to miRNA sequences. In the study by Piriyapongsa

 doi:10.20944/preprints202011.0548.v1

et al. (2007), 462 human miRNA gene sequences from the miRbase database were analyzed, and 68 were shown to contain TE sequences. Further, a negative correlation was observed between the expression level of TE-derived miRNAs and their putative target genes (Piriyapongsa et al., 2007). In another study, miRBase data was analyzed to detect Rdmir (Repeat derived miRNA) in different species, in which a miRNA was defined as a Rdmir if at least 50% of it overlapped with the TE sequence. Using this rule, a total of 226 miRNA genes were identified in humans as Rdmirs (Yuan et al., 2011). Analysis of 6845 pre-miRNAs from eight different vertebrate species in the study by Qin et al., (2015) showed that miRNAs derived from TEs (MDTEs) account for 19.8% of miRNAs in the human genome, which include a total of 409 TE-derived miRNAs (386 overlapped with TEs and 23 un-overlapped with TEs). The proportion was higher than those of other vertebrates. MDTEs with un-overlapped TEs are those miRNAs that are derived from TEs but losing their TE sequences during evolution. Such MDTEs were determined by analyzing miRNAs un-overlapped with TEs and comparing them with homologues in other vertebrates. A total of twenty-three such miRNAs were identified in human that have no TE overlap but their homologues in other species have TE sequences. After excluding multi-copy MDTEs, 338 unique MDTEs (UMDTEs) were identified. These UMDTEs were further classified into type I UMDTEs derived from inverted TE sequences (11.24%), type II UMDTEs with sequences partly overlapping with TE sequences that were not inverted (51.78%), and type III UMDTEs with sequences entirely derived from TE sequences (36.98%) (Qin et al., 2015). A database named MDTE DB (A Database for MicroRNAs Derived from Transposable Element) catalogues all the MDTEs identified by computational analysis of pre-miRNA sequences in miRbase (v20). The database reports 2853 MDTEs. In humans, about 250 partially covered and 150 wholly covered MDTEs have been identified (Wei et al., 2016). It is worth noting that these studies analyzed miRNA sequences from earlier versions

 doi:10.20944/preprints202011.0548.v1

of miRbase. The miRbase archive of miRNA sequences has been increasing quickly and the latest version miRBase (v22) released in 2018 reports 48,860 mature microRNAs from 271 organisms (Kozomara et al., 2019). There are more than 20,000 new entries in this version and the sequence has been changed for more than 800 entries. This demands the latest update of MDTEs based on the current version of miRbase.

Like for miRNAs, the contribution of transposons in human lncRNAs has also been established by several studies. For examples, a study analyzed 19,835 lncRNA transcripts from Gencode v13 and found that 75% of these lncRNAs transcripts have TE sequences (Kapusta et al., 2013). In another study, 61 of the 94 human lncRNA transcripts (65%) in the lncRNA database (lncRNAdb) were shown to have embedded TEs, making 27% of these lncRNA transcript sequences in length. lncRNA genes harboring TEs were enriched in human chromosome 11, while chromosomes 16, 17, and 21 lacked lncRNAs containing TEs (Kang et al., 2015). With the consistent growth of Gencode over different versions, the recent release of Gencode (v34, April 2020) catalogs 17,960 lncRNA genes and 270,000 transcripts (Ramakrishnaiah et al., 2020), justifying an updated study regarding TE-derived sequences in lncRNA genes. Moreover, because of differences in the definitions of what constitutes lncRNA, the number of lncRNAs in the human genome drastically varies across different databases including Gencode (Harrow et al., 2012), FANTOM CAT (Hon et al., 2017), NONCODE (Fang et al., 2018) among others. To address this issue, large scale annotations combining all lncRNA databases into one compendium are provided by the European Bioinformatics Institute (EMBL-EBI) comprehensive database RNACentral (The RNAcentral Consortium et al., 2017). Another highly consistent database is LNCipedia that also provides functional annotations on lncRNA genes by an extensive manual literature curation. So far, 1555 lncRNA genes are annotated with functional information (Volders et al., 2019). Analyzing these

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

all-inclusive lncRNA datasets and functionally annotated lncRNAs for embedded TE sequences might provide a rational extension to the existing studies.

Many lncRNAs are transcribed from intergenic regions (lincRNAs) and play a crucial role in gene regulation. lincRNAs constitute most of the lncRNAs and they are considered as the largest class of ncRNAs in the human genome with >8000 lincRNA genes defined (Cabili et al., 2011). Thus, there are studies explicitly focusing on lincRNAs. The study by Kelly and Rinn (2012) provided a comprehensive analysis of human TE sequences in lincRNAs by obtaining RNA-seq data for 28 different tissues and cell lines. It was found that 7700 lincRNAs overlapped with TEs and 1530 lincRNAs were depleted of TEs, indicating 80% of lincRNA genes associated with TEs and TEs comprised 42% of the total lncRNA sequences (Kelley & Rinn, 2012). In a work conducted by Kannan and coworkers (2015), 69% of 589 human lincRNAs from the NRED database were found to have TE-derived sequences (a fraction lower than 80% determined by (Kelley & Rinn, 2012)). Further, different regions of human lincRNA genes were analyzed for the contribution of TEs. The percentage of TE-derived sequences in lincRNA genes was the highest for introns (>45%), followed by exons (>20%) and promoters (>10%). The distribution was similar to that of protein-coding genes. However, the content of TEs in lincRNA genes was substantially higher than that in protein-coding genes, especially in exons and promoter regions, which is indicative of the low functional constrains for lncRNA genes (Kannan et al., 2015).

TEs have therefore clearly made significant contribution to regulatory RNAs (miRNAs and lncRNAs) sequences. Palindromic sequences of certain TE families play crucial roles for the hairpin structure of miRNAs and different TEs are linked to different miRNA families. TE sequences have also been found in non-hairpin mature miRNAs. The presence of TEs in all regions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

of lncRNA genes (promoters, introns, and exons) highlights TEs' contribution in the generation of regulatory RNAs as well as in the processed and unprocessed forms of these RNAs.

3.2 Functional significance of TEs in regulatory RNA sequences

TE-derived sequences also impart functional properties to different types of sncRNAs and lncRNAs, making them essential for regulatory RNA functions, as demonstrated by the studies described below.

First, the TE-derived sequences have crucial roles in different types of human sncRNAs. miRNAs harboring TE sequences have been found to target genes having embedded TE sequences in 3' UTR. For example, LINE2-derived miR-28-5p and miR-151 target Ly6/Plaur domain-containing 3 (*LYPD3*) and ATP synthase mitochondrial F1 complex assembly factor 1 (*ATPAF1*) genes respectively through pairing to LINE2 elements on 3' UTR (Shin et al., 2010). The subsequent study showed that miR-28-5p also regulates the expression of *LYPD3* and E2F transcription factor 6 (*E2F6*) genes through 3' UTR harboring LINE2 sequences (Spengler et al., 2014).

Second, TEs have also been found to have a diverse role in human lncRNA functions. In a study by Cartault et al. (2012), it was observed that a single point mutation in the L1 of an lncRNA causes a decrease in the level of the lncRNA and is associated with the brain stem cell atrophy. Although the underlying mechanism is poorly understood, the study highlights the functional significance of TE sequences in lncRNAs (Cartault et al., 2012). *Alu* sequences are involved in the base pairing of lncRNA to its target mRNA, which is required for decaying target mRNA. In such cases, *Alu* sequences are present on both lncRNA and mRNA, which can lead to the formation of short imperfect pairing between the two RNA molecules. For example, a 3' UTR *Alu* element of the plasminogen activator inhibitor type 1 (*SERPINE1*) gene binds to lncRNA harboring *Alu*

 doi:10.20944/preprints202011.0548.v1

sequences. The dsRNA structure is further degraded through staufeb1-mediated decay (Gong & Maquat, 2011). *Alu* elements have also been proposed to be involved in the circularization of circular lncRNAs. Circular lncRNAs make an important class of regulatory RNAs and impact gene regulation by influencing the transcription, mRNA turnover, and translation. They harbor exons out of order from the genomic context and are generated by exon shuffling (non-colinear splicing). *Alu* sequences in introns flanking the exons are thought to produce circularization through *Alu*/*Alu* base pairing (Jeck et al., 2013). TEs also provide pre-formed structural and sequence features to lncRNAs, which imparts them the ability to interact with other biological molecules including DNA, RNA, and protein. The RIDL (Repeat Insertion Domain of lncRNA) hypothesis was proposed based on the concept that TEs serve as the functional domain of lncRNA (Johnson & Guigó, 2014). For example, the ERVB5 sequence on XIST lncRNA provides binding sites for polycomb repressive complex 2 (PRC2) that contributes to chromatin compaction (Elisaphenko et al., 2008). TEs have a significant influence on the lncRNA gene structure, and it has been found that TE-derived sites are present in promoters, splice donors, splice acceptors, and polyadenylation sites of lncRNA genes (Kapusta et al., 2013). In a study by Kelley and Rinn (2012), 127 lncRNAs were found to be upregulated by an HERV-H element acting as promoters of these lncRNAs. Based on this observation, it was proposed that TEs, such as HERV-H, can give rise to new lncRNAs by inserting active promoters into previously inactive genomic regions (Kelley & Rinn, 2012). TEs have also been proposed to assist lncRNA in the formation of stable secondary structures. To assess this hypothesis, a study retrieved lncRNA data from GENCODE and compared lncRNAs with TEs to lncRNAs without TEs. Comparing the minimum free energy (MFE) of predicted secondary structures using the program randfold determined that lncRNAs with TEs form more stable secondary structures than those without TEs (Kapusta et al., 2013).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

Another line of evidence supporting the role of TEs in promoting secondary structures in lncRNAs, came from the analysis of A to I editing sites in lncRNAs. Inosine base pairs with cytidine and the editing of adenosine to inosine modulates base pairing of the dsRNA. It was found that about 82% of RNA editing sites locate in the *Alu* regions of lncRNAs. This suggests the *Alu* regions in regulatory RNAs are involved in inter- and intra-molecular base pairing to form stable secondary structures (Kapusta et al., 2013).

In summary, the findings of different studies indicate a clear role of TEs in the functionality of regulatory RNAs in different ways, including, but not limited to, helping the circularization of circular lncRNAs, binding of regulatory RNA to target mRNAs, and formation of the stable secondary structure of regulatory RNAs.

3.3 Role of TEs in lineage specificity of regulatory RNAs

Several studies have reported the lineage specificity of TE-derived regulatory RNAs. For example, the work by Piriyapongsa et al., (2007) which examined the per-site conservation scores of miRNA sequences in the miRbase data, showed that on average, TE-derived miRNAs are less conserved than non-TE-derived miRNAs. Out of 55 TE-derived miRNAs, only 18 were found as conserved (conservation score above a fixed threshold) and 37 were non-conserved. The least-conserved ones were primate-specific (Piriyapongsa et al., 2007). As another example, a placental-specific miRNA gene family mir-1302 has all its members derived from MER53 transposons (eutherian-specific TE) with 58 potential orthologs in placental mammals, indicating the emergence of this miRNA family after the placental mammals diverged from marsupials (Yuan et al., 2010). As shown in another study by Qin et al., the proportions of TE-derived miRNA increased with the evolution of vertebrates from less than 5% in zebrafish to ~20% in humans. Further, sequence analysis of these miRNAs showed no homology among these TE-derived miRNAs from *Danio rerio*, *Gallus gallus*,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

and mammals, indicating that TE-derived miRNAs were species-specific due to species-specific TE transpositions (Qin et al., 2015).

lncRNAs have a significant role in the evolution of key regulatory networks underlying the evolutionary processes (Mattick, 2009). TEs likely have contributed to the functional evolution of lncRNA genes (Johnson & Guigó, 2014). The insertion of TEs in lncRNA genes is considered as an important mechanism behind lineage-specific changes in lncRNAs-mediated gene regulation. Primate-specific TEs were identified in the known TSSs of eight functionally characterized lncRNAs, suggesting the role of TEs in the birth of these lncRNAs during primate evolution (Kapusta et al., 2013). Another study by Kannan et al. determined the evolutionary rate of human lncRNAs by estimating pairwise evolutionary distances for human–macaque alignment and found a significant positive correlation between TE content and the evolutionary rate of lncRNAs (Kannan et al., 2015). As an example, in case of *Xist* lncRNA, many TEs are already present in the *Xist* locus of Eutherian ancestor involved in the generation of the first functional *Xist* transcript. However, many other TEs in the *Xist* exons are lineage-specific and contribute to *Xist*'s functional diversification during Eutherian evolution (Elisaphenko et al., 2008).

In summary, TE-derived regulatory RNAs tend to be less conserved and lineage-specific, implicating TEs as an important source of lineage specificity of regulatory RNAs.

3.4 Tissue-specificity of TE-derived regulatory RNAs

Beyond lineage-specificity, studies have also shown that TE-enriched regulatory RNAs can be tissue-specific. For example, in the study by Kang et al., a total of 29 human lncRNAs were found to have tissue-specific expression, out of which 20 were TE-derived lncRNAs. Moreover, 9 of the 11 lncRNAs found to be expressed in cancer cell lines contained TE sequences, indicating the role

 doi:10.20944/preprints202011.0548.v1

of TE-embedded lncRNAs in cancer (Kang et al., 2015). In another study, it was observed that 127 human lincRNAs having HERV-H sequences were expressed at much higher levels in pluripotent cells, H1-hESCs, and iPSCs, with HERVH LTR in the TSSs of the lncRNA genes, suggesting that TEs might induce tissue-specific expression in these cases (Kelley & Rinn, 2012). The TE-driven tissue-specific expression of lncRNAs has been further elucidated in the study by Chishima et al, which identified many TE–tissue pairs associated with tissue-specific expression of lncRNAs using tissue expression data of human lncRNAs from three different datasets of 'Expression Atlas'. For example, ERV1-lncRNAs were shown to express specifically in testis and L1PA2 was shown to promote the placental specific expression of L1PA2-lncRNAs with the antisense promoter of L1PA2 overlapping with the TSS-neighboring region of lncRNAs being the likely driver of tissue-specific expression (Chishima et al., 2018).

In conclusion, regulatory RNAs with embedded TE sequences have been revealed to have tissue-specific expression patterns. Moreover, investigating the region of overlap between TEs and these RNA sequences in some cases, highlight that TEs in the TSS neighboring region of lncRNAs might be responsible for driving tissue-specific expression.

3.5 Differential contribution to regulatory RNAs among TE types

Different types of TEs have a varying contribution to human regulatory RNA sequences. For miRNAs, the study by Qin et al. classified TE-derived human miRNAs from miRbase in three different types and found 1) SINEs and LINEs are the major contributors to miRNA sequences with inverted TE sequences; 2) SINEs, LINEs, and DNA transposons are major contributors to miRNAs with partial overlaps with non-inverted TE sequences; 3) DNA transposons and SINEs are the primary contributors to miRNA derived entirely from TEs. LTR retrotransposons were thus found to have the least contribution in all three types of miRNAs (Qin et al., 2015).

 doi:10.20944/preprints202011.0548.v1

Several studies also examined the TE composition of human lncRNAs. A study found that SINEs and LINEs as the prevalent TE types contributed 29% of the sequences for the 7700 TE-derived lincRNAs, despite shown as depleted compared to their genome averages (L1s depleted by 2-fold and *Alu* elements depleted by 1.4-fold), while LTR families were showed to be enriched in these lncRNAs despite not being a major TE contributor (Kelley & Rinn, 2012). Kang and coworkers found that 61 of the 94 human lncRNA sequences from lncRNAdb had TEs, most belonging to SINEs and LINEs. The percentage of lncRNA sequence contributed by different types of TEs was 13% for LINEs, 7.7% for SINEs, 3.5% for LTRs, and 2.2% for DNAs, with *Alu*Sx and L1 subfamilies having the highest copy number (Kang et al., 2015). Thus, both of the above studies showed that SINEs and LINEs contribute most to the lncRNA sequences but in less proportion compared to their contribution in the whole genome. This is further supported in the study by Kapusta and coworkers, which in analysis of human lncRNA sequences from Gencode, showed that LINEs were under-represented and LTRs were over-represented in lncRNA sequences (~30% vs. ~40% for LINEs and 30% vs. 20% for LTRs in the lncRNAs vs the genome, respectively). Further, LTRs were over-represented in the exonic and proximal region of lncRNA genes than that of protein-coding genes (Kapusta et al., 2013). In another study, different regions of lincRNA genes (from NRED – Non-encoding RNA expression database) in the human genome were analyzed to assess the contribution of different TE types. It was observed that the distribution of TEs in the introns of lincRNA genes was similar to that in the whole genome, indicating no bias for specific TE type. However, there was a significant reduction of LINEs in exonic and promoter regions of lincRNA genes. Compared to ~20% of the sequence covered by LINEs in the whole genome, sequence coverage of LINEs was ~5% for both lincRNA exonic and promoter regions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

This might be because LINEs have a deleterious impact when inserted into the functional regions of genes (Kannan et al., 2015).

From the findings of the studies mentioned above, concludingly it can be said that among all TEs, SINEs and LINEs contribute most to the lncRNA sequence. However, in contrast to the whole genome, SINEs and LINEs are under-represented, while LTRs are overrepresented in lncRNAs. Regarding different regions of lncRNA genes, comparisons were made with the TEs distribution in whole genome and corresponding regions of protein-coding genes. In summary, TEs' distribution in introns of lncRNA genes is roughly similar to that of the whole genome, but in exonic and promoter regions LINEs are under-represented. Comparison of exonic and promoter regions of lncRNA genes with that of protein-coding genes revealed that LTRs are over-represented in the case of exons and promoters of lncRNAs.

4. Impact of processed pseudogenes on gene regulation

This section briefly discusses the regulatory potential of processed pseudogenes in the human genome. Processed pseudogenes are generated by LINE-mediated retrotransposition of mRNAs. Unlike their parental genes, they lack promoters and introns. Gencode, a large-scale project providing gene annotations in the human genome, has reported 10,668 processed pseudogenes in the human genome representing 72% of all human pseudogenes (Pei et al., 2012). The majority of processed pseudogenes in humans were originated after a split between primates and rodents, corresponding to the period of *Alu* amplification 40-60 million years ago (Ohshima et al., 2003). Formation of processed pseudogenes is ongoing (Maranda et al., 2019) and more than 40 polymorphic processed pseudogenes have been reported in the human genome (Ewing et al., 2013).

 doi:10.20944/preprints202011.0548.v1

The role of processed pseudogenes in gene regulation is an intriguing area to study, and there are several ways, in which they could regulate gene expression. In a study by Harrison et al., a total of 234 transcribed processed pseudogenes in the human genome have been identified by mapping expressed sequences onto processed pseudogenes and identifying identical coding-sequence disablements in both the expressed and genomic sequences (Harrison et al., 2005). Evidence of transcribed processed pseudogenes in the human genome implies that processed pseudogenes might have regulatory potential similar to non-processed pseudogenes that get transcribed and regulate parent gene expression by producing antisense transcripts or by competing for miRNA binding sites with the parent gene's mRNA (Delpu et al., 2016). Processed pseudogenes have also been found to contribute to regulatory RNA sequences. The findings of Milligan et al., (2016) revealed exon-to-exon overlap between processed pseudogenes and lncRNA genes in human genome. Moreover, in the study by Podlaha et al., transcribed processed pseudogene of the *Makorin1* gene in mice has been found to stabilize the transcript of the parental functional gene (Podlaha, 2004), suggesting a similar role for the processed pseudogenes in humans. Another aspect of processed pseudogenes' impact on gene expression is the downregulation of genes due to insertion of processed pseudogene in genic regions. A few such cases have been observed in certain diseases. For example, a processed pseudogene insertion in *CYBB* (cytochrome b-245, beta polypeptide) gene has been associated with chronic granulomatous disease (de Boer et al., 2014), and a processed pseudogene insertion found in tumor suppressor gene, *MGA1* has been linked to lung adenocarcinoma (Cooke et al., 2014).

In conclusion, processed pseudogenes are potentially important candidates for impacting gene regulation. The discovery of transcribed processed pseudogenes and the overlap of processed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 November 2020 doi:10.20944/preprints202011.0548.v1

pseudogenes with the lncRNA genes highlight their regulatory potential for the parental gene. Moreover, by inserting into genic regions, they may downregulate the host genes.

5. Summary and Perspectives

This review considers two different aspects of TEs' contribution to gene regulation: in *cis*-regulatory sequences, and in regulatory RNAs (Figure 2).

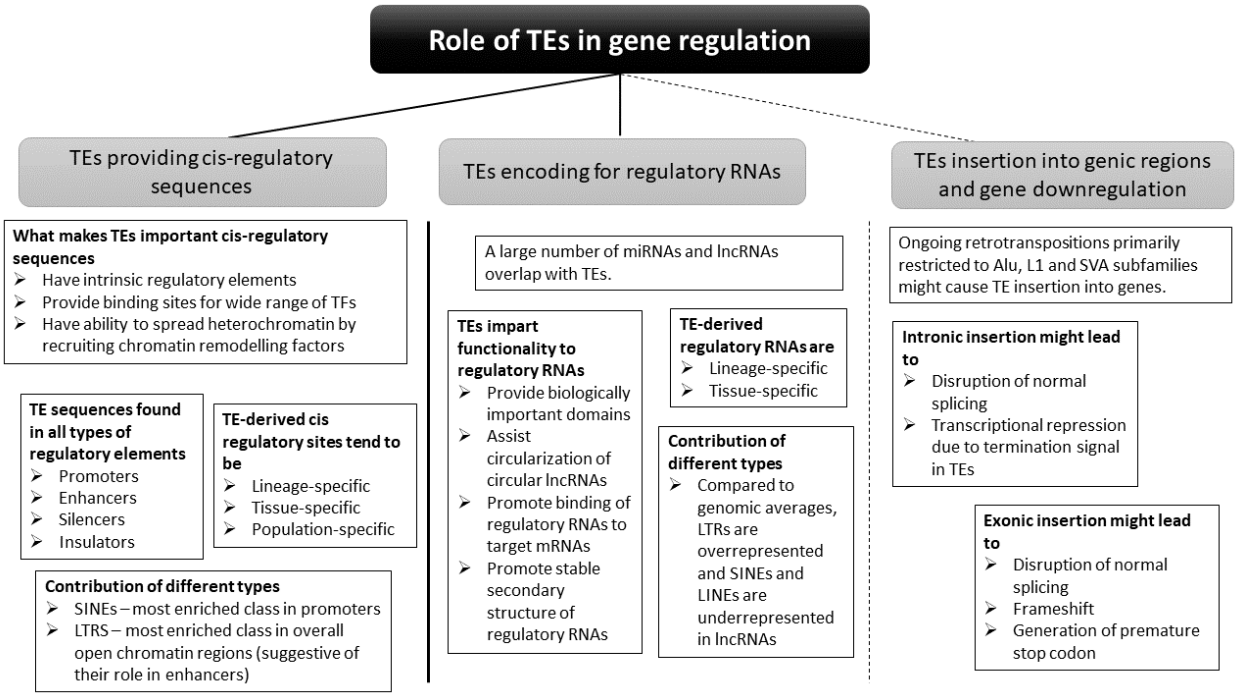

Figure 2. Different ways by which TEs contribute to gene regulation.

TEs have intrinsic regulatory properties for regulating their own expression and provide ready-to-use TFBSs or undergo mutations to provide binding motifs for TFs. TE sequences have been found in the regulatory elements of many genes, participating in short-range and long-range control of gene expression. Among different classes of TEs, SINEs have the highest contribution in all types of regulatory regions. Genes with tissue-specific expression are more likely to have TE sequences in the regulatory regions. TE-derived regulatory sites tend to be lineage-specific as well as species-

 doi:10.20944/preprints202011.0548.v1

specific. Furthermore, polymorphic TEs have been associated with gene expression differences among populations or even individuals.

TEs also contribute to gene regulation by directly participating in the generation of regulatory RNAs. Some TE types are associated explicitly with certain miRNA families. TE sequences in the regulatory RNAs are crucial for their regulatory function by assisting in formation of secondary structures of regulatory RNAs and in binding of regulatory RNAs to their target mRNA sequences. TEs also provide sequence and structural motifs to regulatory RNAs that facilitates the interaction with other biological molecules. Like the TE-derived cis-regulatory sequences, TE-derived regulatory RNA sequences tend to be lineage-specific as well. Furthermore, the tissue-specific expression of TE-derived regulatory RNAs started to be recognized. Among different types of TEs, SINEs and LINEs contribute most to lncRNA sequence, and DNA transposons and SINEs are the major contributors for miRNAs entirely derived from TEs. Processed pseudogenes, a side-products of L1 transposition, when expressed transcribed, can encode lncRNAs and have different ways for regulating the expression of the parent genes.

Research on TEs' role in gene regulation is still in its early stage, leaving ample room for further investigation. For example, systematic studies are needed to comprehensively unveil the contribution of different TE types in the cis-regulatory regions and regulatory RNA sequences using databases providing the most recent annotations. Moreover, there is a need to comprehensively analyze the evolutionary dynamics of these TE-derived regulatory elements genome-wide, instead of focusing on particular subsets. Additionally, there is a need to correlate polymorphisms of TE-derived regulatory elements with the different gene expression patterns among populations and even individuals. Such types of studies demand specialized datasets providing genotype calls of the TEs present in regulatory regions and matching gene expression

data of the same individuals. Experimental verification of the functional impact of TEs on gene regulation is also essential. Analyzing the regulatory potential of regulatory elements with and without TE sequences using the reporter gene expression approach might further support the role of TEs in gene regulation.

Acknowledgments: This work is in part supported by grants from the Canadian Research Chair program, Canadian Foundation of Innovation, Ontario Ministry of Research and Innovation, Canadian Natural Science and Engineering Research Council (NSERC), and Brock University to P.L.

Author contributions: Conceptualization, A.A. and P.L.; Writing – Original Draft Preparation, A.A.; Writing – Review & Editing, A.A, K.H, and P.L.; Supervision, P.L.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ahmed, M., & Liang, P. (2012). Transposable elements are a significant contributor to tandem repeats in the human genome. *Comparative and Functional Genomics*, *2012*, 947089. https://doi.org/10.1155/2012/947089

Ahn, K., Gim, J.-A., Ha, H.-S., Han, K., & Kim, H.-S. (2013). The novel MER transposon-derived miRNAs in human genome. *Gene*, *512*(2), 422–428. https://doi.org/10.1016/j.gene.2012.08.028

Allet, B. (1979). Mu insertion duplicates a 5 base pair sequence at the host inserted site. *Cell*, *16*(1), 123–129. https://doi.org/10.1016/0092-8674(79)90193-4

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J.

A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Ayarpadikannan, S., & Kim, H.-S. (2014). The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics*, *12*(3), 98. https://doi.org/10.5808/gi.2014.12.3.98

Azlan, A., Dzaki, N., & Azzam, G. (2016). Argonaute: The executor of small RNA function. *Journal of Genetics and Genomics*, *43*(8), 481–494. https://doi.org/10.1016/j.jgg.2016.06.002

Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *The American Journal of Human Genetics*, *73*(4), 823–834. https://doi.org/10.1086/378594

Batzer, M. A., & Deininger, P. L. (1991). A human-specific subfamily of Alu sequences. *Genomics*, *9*(3), 481–487. https://doi.org/10.1016/0888-7543(91)90414-A

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, *28*(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

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. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Brattås, P. L., Jönsson, M. E., Fasching, L., Nelander Wahlestedt, J., Shahsavani, M., Falk, R., Falk, A., Jern, P., Parmar, M., & Jakobsson, J. (2017). TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell Reports*, *18*(1), 1–11. https://doi.org/10.1016/j.celrep.2016.12.010

Brini, A. T., Lee, G. M., & Kinet, J. P. (1993). Involvement of Alu sequences in the cell-specific regulation of transcription of the γ chain of Fc and T cell receptors. *Journal of Biological Chemistry*, *268*(2), 1355–1361.

Britten, R. J. (2010). Transposable element insertions have strongly affected human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(46), 19945–19948. https://doi.org/10.1073/pnas.1014330107

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–1927. https://doi.org/10.1101/gad.17446611

Cartault, F., Munier, P., Benko, E., Desguerre, I., Hanein, S., Boddaert, N., Bandiera, S., Vellayoudom, J., Krejbich-Trotot, P., Bintner, M., Hoarau, J. J., Girard, M., Geńin, E., De Lonlay, P., Fourmaintraux, A., Naville, M., Rodriguez, D., Feingold, J., Renouil, M., … Henrion-Caude, A. (2012). Mutation in a primate-conserved retrotransposon reveals a noncoding RNA as a mediator of infantile encephalopathy. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(13), 4980–4985. https://doi.org/10.1073/pnas.1111596109

Chishima, T., Iwakiri, J., & Hamada, M. (2018). Identification of transposable elements contributing to tissue-specific expression of long non-coding RNAs. *Genes*, *9*(1). https://doi.org/10.3390/genes9010023

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Conley, A. B., Piriyapongsa, J., & Jordan, I. K. (2008). Retroviral promoters in the human genome. *Bioinformatics*, *24*(14), 1563–1567. https://doi.org/10.1093/bioinformatics/btn243

Cooke, S. L., Shlien, A., Marshall, J., Pipinikas, C. P., Martincorena, I., Tubio, J. M. C., Li, Y., Menzies, A., Mudie, L., Ramakrishna, M., Yates, L., Davies, H., Bolli, N., Bignell, G. R.,

Tarpey, P. S., Behjati, S., Nik-Zainal, S., Papaemmanuil, E., Teixeira, V. H., … Campbell, P. J. (2014). Processed pseudogenes acquired somatically during cancer development. *Nature Communications*, *5*(1), 3644. https://doi.org/10.1038/ncomms4644

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703. https://doi.org/10.1038/nrg2640

Czipa, E., Schiller, M., Nagy, T., Kontra, L., Steiner, L., Koller, J., Pálné-Szén, O., & Barta, E. (2020). ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them. *Database*, *2020*. https://doi.org/10.1093/database/baz141

de Boer, M., van Leeuwen, K., Geissler, J., Weemaes, C. M., van den Berg, T. K., Kuijpers, T. W., Warris, A., & Roos, D. (2014). Primary Immunodeficiency Caused by an Exonized Retroposed Gene Copy Inserted in the CYBB Gene. *Human Mutation*, *35*(4), 486–496. https://doi.org/10.1002/humu.22519

Deininger, P. L., Moran, J. V, Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development*, *13*(6), 651–658. https://doi.org/10.1016/j.gde.2003.10.013

Delpu, Y., Larrieu, D., Gayral, M., Arvanitis, D., Dufresne, M., Cordelier, P., & Torrisani, J. (2016). Noncoding RNAs. In *Drug Discovery in Cancer Epigenetics* (pp. 305–326). Elsevier. https://doi.org/10.1016/B978-0-12-802208-5.00012-6

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Diehl, A. G., Ouyang, N., & Boyle, A. P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nature Communications*, *11*(1), 1–18. https://doi.org/10.1038/s41467-020-15520-5

Dikstein, R. (2012). Transcription and translation in a package deal: The TISU paradigm. *Gene*,

491(1), 1–4. https://doi.org/10.1016/j.gene.2011.09.013

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108. https://doi.org/10.1038/nature11233

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N., & Zakian, S. M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE*, *3*(6), 1–11. https://doi.org/10.1371/journal.pone.0002521

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, *9*(3), 215–216. https://doi.org/10.1038/nmeth.1906

Ewing, A. D., Ballinger, T. J., Earl, D., Harris, C. C., Ding, L., Wilson, R. K., & Haussler, D. (2013). Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biology*, *14*(3), R22. https://doi.org/10.1186/gb-2013-14-3-r22

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., Sun, L., Zhang, M. Q., Chen, R., & Zhao, Y. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research*, *46*(D1), D308–D314. https://doi.org/10.1093/nar/gkx1107

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A. R. R., Suzuki, H., Hayashizaki, Y., Hume, D. A., … Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, *41*(5), 563–571. https://doi.org/10.1038/ng.368

Franchini, L. F., López-Leal, R., Nasif, S., Beati, P., Gelman, D. M., Low, M. J., De Souza, F. J. S., & Rubinstein, M. (2011). Convergent evolution of two mammalian neuronal enhancers by sequential exaptation of unrelated retroposons. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15270–15275.

https://doi.org/10.1073/pnas.1104997108

Gao, T., & Qian, J. (2019). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, *48*(D1), D58–D64. https://doi.org/10.1093/nar/gkz980

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Gong, C., & Maquat, L. E. (2011). ALUstrious long ncRNAs and their roles in shortening mRNA half-lives. *Cell Cycle*, *10*(12), 1882–1883. https://doi.org/10.4161/cc.10.12.15589

Grindley, N. D. F. (1978). IS1 insertion generates duplication of a nine base pair sequence at its target site. *Cell*, *13*(3), 419–426. https://doi.org/10.1016/0092-8674(78)90316-1

Habegger, L., Sboner, A., Gianoulis, T. A., Rozowsky, J., Agarwal, A., Snyder, M., & Gerstein, M. (2011). RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, *27*(2), 281–283. https://doi.org/10.1093/bioinformatics/btq643

Hadjiargyrou, M., & Delihas, N. (2013). The intertwining of transposable elements and non-coding RNAs. *International Journal of Molecular Sciences*, *14*(7), 13307–13328. https://doi.org/10.3390/ijms140713307

Hambor, J. E., Mennone, J., Coon, M. E., Hanke, J. H., & Kavathas, P. (1993). Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. *Molecular and Cellular Biology*, *13*(11), 7056–7070. https://doi.org/10.1128/mcb.13.11.7056

Han, J. S., Szak, S. T., & Boeke, J. D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*, *429*(6989), 268–274. https://doi.org/10.1038/nature02536

Han, K. (2005). Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Research*, *33*(13), 4040–4052. https://doi.org/10.1093/nar/gki718

Hancks, D. C., & Kazazian, H. H. (2010). SVA retrotransposons: Evolution and genetic instability. *Seminars in Cancer Biology*, *20*(4), 234–245. https://doi.org/10.1016/j.semcancer.2010.04.001

Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Research*, *33*(8), 2374–2383. https://doi.org/10.1093/nar/gki531

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M., & Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, *41*(2), 827–841. https://doi.org/10.1093/nar/gks1284

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., Lizio, M., Kawaji, H., Kasukawa, T., Itoh, M., Burroughs, A. M., Noma, S., Djebali, S., Alam, T., Medvedeva, Y. A., … Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, *543*(7644), 199–204. https://doi.org/10.1038/nature21374

Ito, J., Sugimoto, R., Nakaoka, H., Yamada, S., Kimura, T., Hayano, T., & Inoue, I. (2017). Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genetics*, *13*(7), e1006883. https://doi.org/10.1371/journal.pgen.1006883

Jacques, P. É., Jeyakani, J., & Bourque, G. (2013). The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLoS Genetics*, *9*(5). https://doi.org/10.1371/journal.pgen.1003504

Jalali, S., Gandhi, S., & Scaria, V. (2016). Navigating the dynamic landscape of long noncoding RNA and protein-coding gene annotations in GENCODE. *Human Genomics*, *10*(1), 35. https://doi.org/10.1186/s40246-016-0090-2

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F., & Sharpless, N. E. (2013). Erratum: Circular RNAs are abundant, conserved, and associated with ALU repeats (RNA (156)). *Rna*, *19*(3), 426. https://doi.org/10.1261/rna.035667.112.8

Jiang, J.-C., & Upton, K. R. (2019). Human transposons are an abundant supply of transcription factor binding sites and promoter activities in breast cancer cell lines. *Mobile DNA*, *10*(1), 16. https://doi.org/10.1186/s13100-019-0158-3

Jiang, Y., Qian, F., Bai, X., Liu, Y., Wang, Q., Ai, B., Han, X., Shi, S., Zhang, J., Li, X., Tang, Z., Pan, Q., Wang, Y., Wang, F., & Li, C. (2019). SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Research*, *47*(D1), D235–D243. https://doi.org/10.1093/nar/gky1025

Johnson, R., & Guigó, R. (2014). The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs. *Rna*, *20*(7), 959–976. https://doi.org/10.1261/rna.044560.114

Kang, D., Kim, Y. J., Hong, K., & Han, K. (2015). TE composition of human long noncoding RNAs and their expression patterns in human tissues. *Genes and Genomics*, *37*(1), 87–95. https://doi.org/10.1007/s13258-014-0232-7

Kannan, S., Chernikova, D., Rogozin, I. B., Poliakov, E., Managadze, D., Koonin, E. V., & Milanesi, L. (2015). Transposable element insertions in long intergenic non-coding RNA genes. *Frontiers in Bioengineering and Biotechnology*, *3*(JUN), 1–9. https://doi.org/10.3389/fbioe.2015.00071

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L. A., Bourque, G., Yandell, M., & Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, *9*(4). https://doi.org/10.1371/journal.pgen.1003470

Kazazian, H. H. (2004). Mobile Elements: Drivers of Genome Evolution. *Science*, *303*(5664), 1626–1632. https://doi.org/10.1126/science.1089670

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2007). The genome-wide

determinants of human and chimpanzee microsatellite evolution. *Genome Research*, *18*(1), 30–38. https://doi.org/10.1101/gr.7113408

Kelley, D., & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology*, *13*(11), R107. https://doi.org/10.1186/gb-2012-13-11-r107

Kellner, M., & Makałowski, W. (2019). Transposable elements significantly contributed to the core promoters in the human genome. *Science China Life Sciences*, *62*(4), 489–497. https://doi.org/10.1007/s11427-018-9449-0

Kernaleguen, M., Daviaud, C., Shen, Y., Bonnet, E., Renault, V., Deleuze, J.-F., Mauger, F., & Tost, J. (2018). *Whole-Genome Bisulfite Sequencing for the Analysis of Genome-Wide DNA Methylation and Hydroxymethylation Patterns at Single-Nucleotide Resolution* (pp. 311–349). https://doi.org/10.1007/978-1-4939-7774-1_18

Khoury, G., & Gruss, P. (1983). Enhancer elements. *Cell*, *33*(2), 313–314. https://doi.org/10.1016/0092-8674(83)90410-5

Kim, D. S., & Hahn, Y. (2011). Identification of human-specific transcript variants induced by DNA insertions in the human genome. *Bioinformatics*, *27*(1), 14–21. https://doi.org/10.1093/bioinformatics/btq612

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207–220. https://doi.org/10.1038/s41576-018-0089-8

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). MiRBase: From microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. https://doi.org/10.1093/nar/gky1141

Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., Ng, H. H., & Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, *42*(7), 631–634. https://doi.org/10.1038/ng.600

Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M. A., Richter, J., Soler, E., Stadhouders, R., Jöhrens, K., Wurster, K. D., Callen, D. F., Harte, M. F., Giefing, M., Barlow, R., Stein, H., Anagnostopoulos, I., … Mathas, S. (2010).

Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine*, *16*(5), 571–579. https://doi.org/10.1038/nm.2129

Lappalainen, T., Sammeth, M., Friedländer, M. R., ‘t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. https://doi.org/10.1038/nature12531

Liu, N., Lee, C. H., Swigut, T., Grow, E., Gu, B., Bassik, M. C., & Wysocka, J. (2018). Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature*, *553*(7687), 228–232. https://doi.org/10.1038/nature25179

Lynch, V. J., Leclerc, R. D., May, G., & Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, *43*(11), 1154–1159. https://doi.org/10.1038/ng.917

Maranda, V., Sunstrum, F. G., & Drouin, G. (2019). Both male and female gamete generating cells produce processed pseudogenes in the human genome. *Gene*, *684*, 70–75. https://doi.org/10.1016/j.gene.2018.10.061

Marnetto, D., Mantica, F., Molineris, I., Grassi, E., Pesando, I., & Provero, P. (2018). Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *American Journal of Human Genetics*, *102*(2), 207–218. https://doi.org/10.1016/j.ajhg.2017.12.014

Martens, J. H. A., & Stunnenberg, H. G. (2013). BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, *98*(10), 1487–1489. https://doi.org/10.3324/haematol.2013.094243

Mattick, J. S. (2009). Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. *Annals of the New York Academy of Sciences*, *1178*(1), 29–46. https://doi.org/10.1111/j.1749-6632.2009.04991.x

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L., Liu, T., Brown, M., Meyer, C. A., & Liu, X. S. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*,

45(D1), D658–D662. https://doi.org/10.1093/nar/gkw983

Milligan, M. J., Harvey, E., Yu, A., Morgan, A. L., Smith, D. L., Zhang, E., Berengut, J., Sivananthan, J., Subramaniam, R., Skoric, A., Collins, S., Damski, C., Morris, K. V., & Lipovich, L. (2016). Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome. *Frontiers in Genetics*, *7*. https://doi.org/10.3389/fgene.2016.00026

Neznanov, N., Thorey, I. S., Ceceña, G., & Oshima, R. G. (1993). Transcriptional insulation of the human keratin 18 gene in transgenic mice. *Molecular and Cellular Biology*, *13*(4), 2214–2223. https://doi.org/10.1128/mcb.13.4.2214

Nikitin, D., Garazha, A., Sorokin, M., Penzar, D., Tkachev, V., Markov, A., Gaifullin, N., Borger, P., Poltorak, A., & Buzdin, A. (2019). Retroelement-Linked Transcription Factor Binding Patterns Point to Quickly Developing Molecular Pathways in Human Evolution. *Cells*, *8*(2). https://doi.org/10.3390/cells8020130

Nikitin, D., Penzar, D., Garazha, A., Sorokin, M., Tkachev, V., Borisov, N., Poltorak, A., Prassolov, V., & Buzdin, A. A. (2018). Profiling of human molecular pathways affected by retrotransposons at the level of regulation by transcription factor proteins. *Frontiers in Immunology*, *9*(JAN), 1–14. https://doi.org/10.3389/fimmu.2018.00030

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, *4*(11), 1–14. https://doi.org/10.1186/gb-2003-4-11-r74

Ono, M., Kawakami, M., & Takezawa, T. (1987). A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Research*, *15*(21), 8725–8737. https://doi.org/10.1093/nar/15.21.8725

Pace, J. K., & Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Research*, *17*(4), 422–432. https://doi.org/10.1101/gr.5826307

Pavlicev, M., Hiratsuka, K., Swaggart, K. A., Dunn, C., & Muglia, L. (2015). Detecting

endogenous retrovirus-driven tissue-specific gene transcription. *Genome Biology and Evolution*, *7*(4), 1082–1097. https://doi.org/10.1093/gbe/evv049

Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V., & Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. *Nature Communications*, *10*(1), 1–16. https://doi.org/10.1038/s41467-019-13555-x

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J., & Gerstein, M. B. (2012). The GENCODE pseudogene resource. *Genome Biology*, *13*(9), R51. https://doi.org/10.1186/gb-2012-13-9-r51

Peters, L., & Meister, G. (2007). Argonaute Proteins: Mediators of RNA Silencing. *Molecular Cell*, *26*(5), 611–623. https://doi.org/10.1016/j.molcel.2007.05.001

Petri, R., Brattås, P. L., Sharma, Y., Jonsson, M. E., Pircs, K., Bengzon, J., & Jakobsson, J. (2019). LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLoS Genetics*, *15*(3), 1–18. https://doi.org/10.1371/journal.pgen.1008036

Piriyapongsa, J., & Jordan, I. K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*, *2*(2). https://doi.org/10.1371/journal.pone.0000203

Piriyapongsa, J., Mariño-Ramírez, L., & Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, *176*(2), 1323–1337. https://doi.org/10.1534/genetics.107.072553

Podlaha, O. (2004). Nonneutral Evolution of the Transcribed Pseudogene Makorin1-p1 in Mice. *Molecular Biology and Evolution*, *21*(12), 2202–2209. https://doi.org/10.1093/molbev/msh230

Polavarapu, N., Mariño-Ramírez, L., Landsman, D., McDonald, J. F., & King, I. K. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics*, *9*, 1–10. https://doi.org/10.1186/1471-2164-9-226

Qin, S., Jin, P., Zhou, X., Chen, L., & Ma, F. (2015). The role of transposable elements in the origin and evolution of microRNAs in human. *PLoS ONE*, *10*(6), 1–10. https://doi.org/10.1371/journal.pone.0131365

Ramakrishnaiah, Y., Kuhlmann, L., & Tyagi, S. (2020). *Computational approaches to functionally annotate long noncoding RNA (lncRNA). June*. https://doi.org/10.20944/preprints202006.0116.v1

Raviram, R., Rocha, P. P., Luo, V. M., Swanzey, E., Miraldi, E. R., Chuong, E. B., Feschotte, C., Bonneau, R., & Skok, J. A. (2018). Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biology*, *19*(1), 1–19. https://doi.org/10.1186/s13059-018-1598-7

Rayan, N. A., del Rosario, R. C. H., & Prabhakar, S. (2016). Massive contribution of transposable elements to mammalian regulatory sequences. *Seminars in Cell and Developmental Biology*, *57*, 51–56. https://doi.org/10.1016/j.semcdb.2016.05.004

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., & Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. https://doi.org/10.1038/nmeth1068

Roy, A. M., West, N. C., Rao, A., Adhikari, P., Alemán, C., Barnes, A. P., & Deininger, P. L. (2000). Upstream flanking sequences and transcription of SINEs. *Journal of Molecular Biology*, *302*(1), 17–25. https://doi.org/10.1006/jmbi.2000.4027

Samuelson, L. C., Wiebauer, K., Snow, C. M., & Meisler, M. H. (1990). Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Molecular and Cellular Biology*, *10*(6), 2513–2520. https://doi.org/10.1128/mcb.10.6.2513

Sen, S. K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P. A., Dyer, M., Cordaux, R., Liang, P., & Batzer, M. A. (2006). Human Genomic Deletions Mediated by Recombination between Alu Elements. *The American Journal of Human Genetics*, *79*(1), 41–53. https://doi.org/10.1086/504600

Sha, M., Lee, X., Li, X. ping, Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., Keith, J. C., & McCoy, J. M. (2000). Syncytin is a captive retroviral

envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785–789. https://doi.org/10.1038/35001608

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell*, *38*(6), 789–802. https://doi.org/10.1016/j.molcel.2010.06.005

Shooshtari, P., Feng, S., Nelakuditi, V., Foong, J., Brudno, M., & Cotsapas, C. (2018). OCHROdb: a comprehensive, quality checked database of open chromatin regions from sequencing data. *BioRxiv*, 484840. https://doi.org/10.1101/484840

Smit, A. F. A., & Riggs, A. D. (1996). Tiggers and other DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(4), 1443–1448. https://doi.org/10.1073/pnas.93.4.1443

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384. https://doi.org/10.1101/pdb.prot5384

Spengler, R. M., Oakley, C. K., & Davidson, B. L. (2014). Functional microRNAs and target sites are created by lineage-specific transposition. *Human Molecular Genetics*, *23*(7), 1783–1793. https://doi.org/10.1093/hmg/ddt569

Spirito, G., Mangoni, D., Sanges, R., & Gustincich, S. (2019). Impact of polymorphic transposable elements on transcription in lymphoblastoid cell lines from public data. *BMC Bioinformatics*, *20*(Suppl 9), 1–13. https://doi.org/10.1186/s12859-019-3113-x

Stacey, S. N., Kehr, B., Gudmundsson, J., Zink, F., Jonasdottir, A., Gudjonsson, S. A., Sigurdsson, A., Halldorsson, B. V., Agnarsson, B. A., Benediktsdottir, K. R., Aben, K. K. H., Vermeulen, S. H., Cremers, R. G., Panadero, A., Helfand, B. T., Cooper, P. R., Donovan, J. L., Hamdy, F. C., Jinga, V., … Stefansson, K. (2016). Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Human Molecular Genetics*, *25*(5), 1008–1018. https://doi.org/10.1093/hmg/ddv622

Stewart, C., Kural, D., Strömberg, M. P., Walker, J. A., Konkel, M. K., Stütz, A. M., Urban, A. E., Grubert, F., Lam, H. Y. K., Lee, W.-P., Busby, M., Indap, A. R., Garrison, E., Huff, C., Xing,

J., Snyder, M. P., Jorde, L. B., Batzer, M. A., Korbel, J. O., … 1000 Genomes Project. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genetics*, *7*(8), e1002236. https://doi.org/10.1371/journal.pgen.1002236

Sultana, T., Zamborlini, A., Cristofari, G., & Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics*, *18*(5), 292–308. https://doi.org/10.1038/nrg.2017.7

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M. P., & Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research*, *24*(12), 1963–1976. https://doi.org/10.1101/gr.168872.113

Sundaram, V., & Wysocka, J. (2020). Transposable elements as a potent source of diverse cis - regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1795), 20190347. https://doi.org/10.1098/rstb.2019.0347

Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and Cellular Biology*, *10*(12), 6718–6729. https://doi.org/10.1128/MCB.10.12.6718

Tang, W., Mun, S., Joshi, A., Han, K., & Liang, P. (2018). Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Research*, *25*(5), 521–533.

Testori, A., Caizzi, L., Cutrupi, S., Friard, O., De Bortoli, M., Cora', D., & Caselle, M. (2012). The role of Transposable Elements in shaping the combinatorial interaction of Transcription Factors. *BMC Genomics*, *13*(1), 400. https://doi.org/10.1186/1471-2164-13-400

The ENCODE (ENCyclopedia Of DNA Elements) Project. (2004). *Science*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

The RNAcentral Consortium, Petrov, A. I., Kay, S. J. E., Kalvari, I., Howe, K. L., Gray, K. A., Bruford, E. A., Kersey, P. J., Cochrane, G., Finn, R. D., Bateman, A., Kozomara, A., Griffiths-Jones, S., Frankish, A., Zwieb, C. W., Lau, B. Y., Williams, K. P., Chan, P. P., Lowe, T. M., … Dinger, M. E. (2017). RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research*, *45*(D1), D128–D134. https://doi.org/10.1093/nar/gkw1008

Thornburg, B. G., Gotea, V., & Makałowski, W. (2006). Transposable elements as a significant source of transcription regulating signals. *Gene*, *365*(1-2 SPEC. ISS.), 104–110. https://doi.org/10.1016/j.gene.2005.09.036

Trizzino, M., Kapusta, A., & Brown, C. D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics*, *19*(1), 1–12. https://doi.org/10.1186/s12864-018-4850-3

Trizzino, M., Park, Y. S., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G. H., Lynch, V. J., & Brown, C. D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. *Genome Research*, *27*(10), 1623–1633. https://doi.org/10.1101/gr.218149.116

van Regenmortel, M. H., & Mahy, B. W. (Eds.). (2010). *Desk encyclopedia of general virology.* Academic Press.

Vidaud, D., Vidaud, M., Bahnak, B. R., Siguret, V., Gispert Sanchez, S., Laurian, Y., Meyer, D., Goossens, M., & Lavergne, J. M. (1993). Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *European Journal of Human Genetics : EJHG*, *1*(1), 30–36. https://doi.org/10.1159/000472385

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., & Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Research*, *47*(D1), D135–D139. https://doi.org/10.1093/nar/gky1031

Wang, H., Xing, J., Grover, D., Hedges Kyudong Han, D. J., Walker, J. A., & Batzer, M. A. (2005). SVA elements: A hominid-specific retroposon family. *Journal of Molecular Biology*, *354*(4), 994–1007. https://doi.org/10.1016/j.jmb.2005.09.085

Wang, L., Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2017). Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Research*, *45*(5), 2318–2328. https://doi.org/10.1093/nar/gkw1286

Wang, T., Zeng, J., Lowe, C. B., Sellers, R. G., Salama, S. R., Yang, M., Burgess, S. M., Brachmann, R. K., & Haussler, D. (2007). Species-specific endogenous retroviruses shape the

transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(47), 18613–18618. https://doi.org/10.1073/pnas.0703637104

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Watkins, W. S., Feusier, J. E., Thomas, J., Goubert, C., Mallick, S., & Jorde, L. B. (2020). The Simons Genome Diversity Project: A Global Analysis of Mobile Element Diversity. *Genome Biology and Evolution*, *12*(6), 779–794. https://doi.org/10.1093/gbe/evaa086

Wei, G., Qin, S., Li, W., Chen, L., & Ma, F. (2016). MDTE DB: A Database for MicroRNAs Derived from Transposable Element. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(6), 1155–1160. https://doi.org/10.1109/TCBB.2015.2511767

Wu, J., Grindlay, G. J., Bushel, P., Mendelsohn, L., & Allan, M. (1990). Negative regulation of the human epsilon-globin gene by transcriptional interference: role of an Alu repetitive element. *Molecular and Cellular Biology*, *10*(3), 1209–1216. https://doi.org/10.1128/mcb.10.3.1209

Yuan, Z., Sun, X., Jiang, D., Ding, Y., Lu, Z., Gong, L., Liu, H., & Xie, J. (2010). Origin and evolution of a placental-specific microRNA family in the human genome. *BMC Evolutionary Biology*, *10*(1), 1–12. https://doi.org/10.1186/1471-2148-10-346

Yuan, Z., Sun, X., Liu, H., & Xie, J. (2011). MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0017666

Zemojtel, T., Kiebasa, S. M., Arndt, P. F., Behrens, S., Bourque, G., & Vingron, M. (2011). CpG deamination creates transcription factor-binding sites with high efficiency. *Genome Biology and Evolution*, *3*(1), 1304–1311. https://doi.org/10.1093/gbe/evr107

Zeng, L., Pederson, S. M., Cao, D., Qu, Z., Hu, Z., Adelson, D. L., & Wei, C. (2018). Genome-Wide Analysis of the Association of Transposable Elements with Gene Regulation Suggests that Alu Elements Have the Largest Overall Regulatory Impact. *Journal of Computational Biology*, *25*(6), 551–562. https://doi.org/10.1089/cmb.2017.0228

 doi:10.20944/preprints202011.0548.v1