Gene 518 (2013) 231–241

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Review

Retroelements in human disease

Kristel Kaer, Mart Speek *

Department of Gene Technology, Tallinn University of Technology, Akadeemia tee 15, Tallinn 12618, Estonia

ARTICLE INFO

Article history:
Accepted 2 January 2013
Available online 17 January 2013

Keywords:
L1
Alu
SVA
Exonization
Transcriptional interference
Cryptic splice sites

ABSTRACT

Retroelements are an abundant class of noncoding DNAs present in about half of the human genome. Among them, L1, Alu and SVA are currently active. They "jump" by retrotransposition, shuffle genomic regions by 5′ and 3′ transduction, and promote or inhibit gene transcription by providing alternative promoters or generating antisense and/or regulatory noncoding RNAs. Recent data also suggest that retroelement insertions into exons and introns of genes induce different types of genetic disease, including cancer. Retroelements interfere with the expression of genes by inducing alternative splicing via exon skipping and exonization using cryptic splice sites, and by providing polyadenylation signals. Here we summarize our current understanding of the molecular mechanisms of retroelement-induced mutagenesis which causes fifty different types of human disease. We categorize these mutagenic effects according to eleven different mechanisms and show that most of them may be explained either by traditional exon definition or transcriptional interference, a previously unrecognized molecular mechanism. In summary, this review gives an overview of retroelement insertions in genes that cause significant changes in their transcription and cotranscriptional splicing and show a remarkable level of complexity.

Contents

1. Introduction

The human genome contains three major types of actively transposing repeated DNAs known as retroelements (REs) (Ostertag and Kazazian, 2001). These are termed Long Interspersed Nuclear Element (LINE or L1), Alu and SVA repeats. Altogether these REs occupy about one-third of our genome. In terms of the number of copies (cps) and total mass (%) of the genome, they contribute as follows: L1—500,000 cps (17%), Alu—1,100,000 cps (11%) and SVA—3000 cps (0.2%) (Lander et al., 2001). L1 is the only currently active and autonomous retrotransposon which contains an internal RNA polymerase (pol) II promoter (Swergold, 1990) and encodes an RNA binding protein (Hohjoh and Singer, 1996), an endonuclease (Feng et al., 1996) and a reverse transcriptase (Mathias et al., 1991) (Fig. 1A). All these activities are required for L1 retrotransposition by a copy-and-paste mechanism whereby mRNA is first transcribed from genomic L1, then reverse transcribed and inserted back into the genome. Autonomous and retrotransposition-competent L1 is also involved in the mobilization of non-autonomous or passive REs, such as Alu, derived from 7SL RNA (part of the signal recognition particle) (Ullu and

Abbreviations: cps, copies; ORF, open reading frame; pol, polymerase; REs, retroelements; ss, splice site; TI, transcriptional interference.

* Corresponding author at: Department of Gene Technology, Akadeemia tee 15, rm 129, Tallinn University of Technology, Tallinn 19086, Estonia. Tel.: +372 620 4449; fax: +372 620 4400.
E-mail addresses: kristel.kaer@mail.ee (K. Kaer), mart.speek@ttu.ee (M. Speek).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.01.008

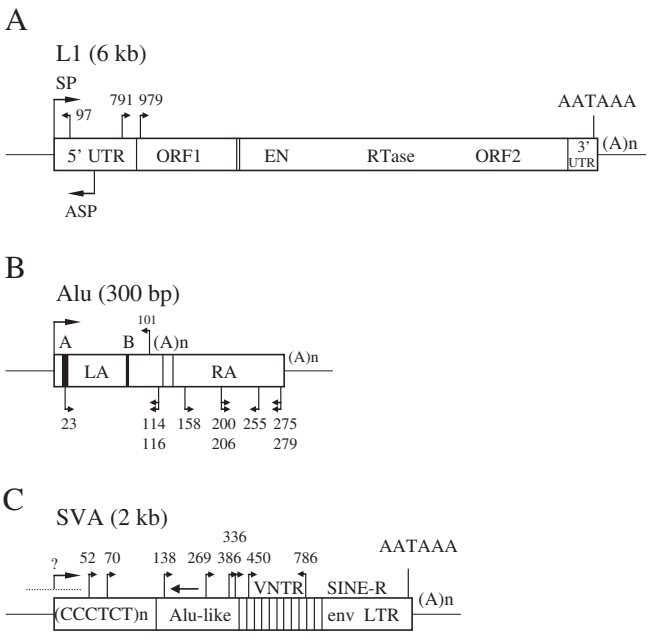

Fig. 1. Canonical structures of three active human REs. (A) A full-length L1 contains sense (SP) (Swergold, 1990) and antisense promoters (ASP) (Speek, 2001) in the 5′ untranslated region (5′ UTR), open reading frames (ORF1) and (ORF2) and 3′ UTR encompassing polyadenylation signal (AATAAA) and polyA tail (A)n. ORF2 encodes two enzymatic activities: endonuclease (EN) (Feng et al., 1996) and reverse transcriptase (RTase) (Mathias et al., 1991). While SP is required for L1 transcription and retrotransposition, ASP drives transcription in opposite direction into adjacent cellular genes producing chimeric transcripts. For both promoters direction of transcription is shown by arrow. (B) A full-length Alu derived from 7SL RNA after imperfect dimerization yielding left (LA) and right (RA) arms separated and terminating with polyA sequence (A)n. Transcription of Alu is carried out by RNA pol III (direction shown by arrow) from the internal promoter, boxes A and B located at positions 4–37 and 70–86, respectively (Grover et al., 2005). (C) A full-length SVA (SINE-VNTR-Alu) contains CCCTCT repeats, Alu-like region in reverse orientation compared to Alu in (A), variable number of GC-rich tandem repeats (VNTR) and part of the envelope gene (env) and right long terminal repeat (LTR) derived from HERV-K10 (SINE-R) (Hancks and Kazazian, 2010). SVA is transcribed by RNA pol II, but the location of its promoter is not known. Transcription of genomic SVAs may be initiated from upstream or within SVA sequence (marked with ?). Small arrows pointing to the left and right show 5′ and 3′ss located in sense (top) and opposite in antisense (bottom) strand in L1 (Belancio et al., 2006), Alu (Sorek et al., 2002) and SVA (Hancks et al., 2009). Only splice sites used at least two (L1 and SVA) and three or more times (Alu) are shown. Approximate sizes of full-length REs are given in parenthesis for each element. Their depicted structures are not drawn to scale.

Tschudi, 1984), and the composite SVA elements, derived from different repeats (Hancks et al., 2011). Both Alu and SVA lack activities necessary for independent transposition (Figs. 1BC). Each human being has about 100 cps of transposition competent L1s, some of which are highly active, as revealed by test reactions (Beck et al., 2010; Brouha et al., 2003).

Most of the REs, interspersed nearly equally between intergenic and intragenic regions of human chromosomes, have become homozygous. Thus, they have been fixed in the human population. Nevertheless, recent genome-wide analysis revealed thousands of polymorphic REs, i.e., they are present or absent in certain gene alleles or genomic loci (Ewing and Kazazian, 2010; Huang et al., 2010; Stewart et al., 2011; Xing et al., 2009). Using next-generation sequencing, Stewart et al. (2011) presented a comprehensive map of 7380 RE (L1, Alu and SVA) polymorphisms derived from 185 samples from European, African and Asian population groups. Based on the number of insertions and estimated time difference from the common ancestor, Xing et al. (2009) calculated retrotransposition frequencies per birth of 1/108, 1/21 and 1/916 for L1, Alu and SVA, respectively. These data show that besides fixed retrotransposons, individual-specific variation of these elements exists, which together with single nucleotide polymorphism (SNP) and copy number variation (CNV) could influence an individual's phenotype and predisposition to disease.

L1 retrotransposition occurs mainly during early embryogenesis (Kano et al., 2009; van den Hurk et al., 2007). As a source of active L1 retrotransposons, L1 mRNA packaged into a viral-like ribonucleoprotein particle may be present in both male and female germ cells and carried into the zygote. After reverse transcription, genomic insertion of L1 cDNA can occur in preimplantation embryos (Singer et al., 2010). However, recent data from several laboratories suggest that L1 retrotransposition or L1-mediated insertion of Alus and SVAs may also occur in somatic cells, including brain (Baillie et al., 2011; Coufal et al., 2009). Using quantitative multiplex polymerase chain reaction, Coufal et al. (2009) have demonstrated a statistically significant increase in transposed L1 copies in the brain of three individuals when compared to the heart and liver samples of the same individuals. These authors also noted a certain variability between individuals and different areas of the brain. In another study (Baillie et al., 2011), a high-throughput protocol, called retrotransposon capture sequencing, was used to identify numerous L1, Alu and SVA germline insertions, as well as putative somatic insertions in the hippocampus and caudate nucleus of three individuals. These authors identified 7743, 13,692 and 1350 somatic L1, Alu and SVA insertions, respectively. Both these studies (Baillie et al., 2011; Coufal et al., 2009), found that insertions of REs occurred frequently in the exons and introns, or in the vicinity of neuronally expressed genes. However, as in both instances the genomic DNA was pooled from the tissue or a large number of cells, it was suggested that the results can only be interpreted as estimates of retrotransposition and the ultimate proof for retrotransposition should come from the single cell analysis. Surprisingly, a recent genome-wide profiling of 300 single neurons derived from two brain areas (cerebral cortex and caudate nucleus) of three individuals revealed de novo somatic L1 insertion in the range of 0.04 to 0.6 per neuron, corresponding to a minimum of 1 insertion per 25 neurons and a maximum slightly higher than 1 insertion in 2 neurons (Evrony et al., 2012). This frequency is substantially lower than the previous estimate of Coufal et al. (2009) (80 insertions per single cell). The large difference between the results of the above-mentioned studies was explained by the use of direct and less artifact-generating single cell sequencing method. Although Evrony et al. (2012) suggested that L1 retrotransposition in human cortex and caudate is rare, it is possible that higher rates in other brain regions (e.g. hippocampus) and individual-specific differences in the numbers and the activities of "hot" L1s (Beck et al., 2010) could contribute to the variability in retrotransposition rates among individuals. Therefore, it remains to be determined whether RE insertions could contribute to the neuron-to-neuron variation and influence the neuronal gene expression either positively, by increasing the diversity of behavioral phenotypes or negatively, by increasing the risk of neurological disorder.

It is known that different environmental effects (ionizing radiation, heavy metals, air pollution, etc.) could trigger the activation of REs (Kale et al., 2005; Tanaka et al., 2012), although no direct relationship between the disease and their activation has been shown. Generally, REs could change the normal gene expression by affecting genome integrity through recombination involving insertions, deletions, and rearrangements, and/or transcriptional effects, by providing alternative promoters, influencing transcriptional elongation and pre-mRNA splicing. In recent years several excellent reviews have been published that address the influence of RE on genome (in)stability and major aspects of retrotransposition (when and where retrotransposition occurs and how it is regulated) in connection with human disease (Beck et al., 2011; Belancio et al., 2008; Hancks and Kazazian, 2012). However, less attention has been paid to the transcription and splicing mechanisms. In this review we focus on the largest category of transcriptional effects, namely transcriptional and cotranscriptional splicing misregulation caused by insertions or mutations of three currently active non-LTR