Available online at www.sciencedirect.com
ScienceDirect
Gene 396 (2007) 303–311
GENE
SECTION
FUNCTIONAL GENOMICS
www.elsevier.com/locate/gene

Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues

Emmanuelle Lerat *, Marie Sémon

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France

Received 1 December 2006; received in revised form 16 March 2007; accepted 2 April 2007
Available online 6 April 2007
Received by I. King Jordan

Abstract

Transposable elements (TEs) are genomic sequences able to replicate themselves, and to move from one chromosomal position to another within the genome. Many TEs contain their own regulatory regions, which means that they may influence the expression of neighboring genes. TEs may also be activated and transcribed in various cancers. We therefore tested whether gene expression in normal and tumor tissues is influenced by the neighboring TEs. To do this, we associated all human genes to the nearest TEs. We analyzed the expression of these genes in normal and tumor tissues using SAGE and EST data, and related this to the presence and type of TEs in their vicinity. We confirmed that TEs tend to be located in antisense orientation relative to their hosting genes. We found that the average number of tissues where a gene is expressed varies depending on the type of TEs located near the gene, and that the difference in expression level between normal and tumor tissues is greatest for genes that host SINE elements. This deregulation increases with the number of SINE copies in the gene vicinity. This suggests that SINE elements might contribute to the cascade of gene deregulation in cancer cells.

Keywords: Transposable elements; Cancer; Gene expression

1. Introduction

Transposable elements (TEs) have been identified in the genomes of almost all living organisms. They are distinguished by their ability to move from one chromosomal position to another within the genome, and to replicate themselves. Mammalian genomes are particularly TE-rich, and a high proportion of their sequence is derived from TEs (45% of the human genome (International Human Genome Sequencing Consortium, 2001) and 38.5% of the mouse genome (Mouse Genome Sequencing Consortium, 2002)). Another large fraction of mammalian genomes is probably also TE-derived, but has diverged beyond recognition (Medstrand et al., 2005).

TEs are divided into two main classes on the basis of their transposition intermediate (Finnegan, 1992). The first class includes the retrotransposons that use an RNA intermediate, and move by a "copy and paste" mechanism. Within this class, two subclasses have been identified, depending on the presence or absence of a Long Terminal Repeat (LTR) at the extremities of the elements. Non-LTR retrotransposons include two categories of elements: the Long Interspersed Nuclear Elements (LINEs) and the Short Interspersed Nuclear Elements (SINEs), the latter being dependent upon the former to transpose. The second class of TEs consists of the transposons that use a DNA intermediate and move via a "cut and paste" mechanism. With the exception of SINEs, active TEs all contain Open Reading Frames (ORFs) that encode all the proteins required for their transposition. All TEs are dependent on internal regulatory regions. In LTR retrotransposons, the promoter is located within the LTR, whereas in LINEs, it is located in the 5′ UnTranslated Region

Abbreviations: TE, Transposable Element; ORF, Open Reading Frame; LINE, Long Interspersed Nuclear Element; SINE, Short Interspersed Nuclear Element; LTR, Long Terminal Repeat; EST, Expressed Sequence Tag; SAGE, Serial Analysis of Gene Expression; endogenous retrovirus, ERV.

* Corresponding author. Laboratoire Biométrie et Biologie Evolutive, Université Claude Bernard, Lyon 1, UMR–CNRS 5558, Bat. Mendel, 69622 Villeurbanne cedex, France. Tel.: +33 4 72 43 29 18; fax: +33 4 72 43 13 88.
E-mail address: lerat@biomserv.univ-lyon1.fr (E. Lerat).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.04.002

(UTR). Interestingly, one particular human LINE (L1) has a second promoter in the antisense orientation in the +400 to +600 bp region of the 5′ UTR (Nigumann et al., 2002). LTR and LINE elements both use an RNA polymerase II, whereas SINEs have an internal type-III polymerase promoter.

TEs contain internal regulatory regions, and this allows them to influence the transcription of the genes near where they are inserted (Britten and Davidson, 1969). In the mouse genome, some LTR retrotransposons have been reported to act as alternative promoters. They provide exonic regions for a subset of genes expressed in oocytes and preimplantation embryos, and permit synchronous developmental expression of these genes (Peaston et al., 2004). The non-LTR retrotransposon L1 can undergo somatic transposition in mouse brain, which alters the levels of gene expression *in vitro*, potentially influencing neural cell fate and generating neuronal diversity (Muotri et al., 2005). Other examples have shown that ancient insertions have contributed to gene regulation (reviewed in Britten, 1996). For instance, the androgen dependence of a mouse sex-linked gene results from the insertion of a retrotransposon in the 5′ region (Stavenhagen and Robins, 1988). On the genomic scale, various analyses have demonstrated that a high proportion of current promoters is derived from TE sequences, suggesting that TEs may commonly affect the evolution of human gene regulation. For example, one quarter of the experimentally-identified promoters in mammals contain TE-derived sequences (Jordan et al., 2003; Van de Lagemaat et al., 2003). Moreover, some TE insertions have been shown to modify the tissue-specific expression of genes. This is the case for an estrogen biosynthesis gene, *CYP19*, the placenta-specific transcription of which is driven by an alternative promoter derived from an LTR (Van de Lagemaat et al., 2003).

TEs are powerful cause of point mutations and can induce diseases. For example, more than 30 genetic diseases and 16 cancers have been attributed to homologous recombination between SINEs (Deininger and Batzer, 1999). Recombination involving LINE elements has also been observed to lead to the formation of tumors in the esophagus and in the female reproductive tract (Segal et al., 1999). The insertion of a retrovirus at particular locations in the genome can also activate oncogenes (Barklis et al., 1986). For instance, the avian leukosis virus, which induces tumors in chicken, is inserted near the c-myc oncogene (Hayward et al., 1981). This oncogene has also been altered by an intronic insertion of a LINE element in a breast cancer (Morse et al., 1988), by upstream insertions of a LINE in tumor development in dogs (Katzir et al., 1987), and by recombination with a LINE in rat immunocytomas (Pear et al., 1988). Several carcinogenic factors, such as benzo(a)pyrene (Stribinskis and Ramos, 2006) or gamma radiation (Farkash et al., 2006), have also been shown to activate TEs. All these findings imply that the epigenetic activation of TEs is able to cause point mutations and genomic instability, thus leading to cancer.

Some studies have shown that TEs can be activated under tumor conditions, although no direct relationship has been demonstrated between the disease and the element. For example, some specific endogenous retroviruses (ERVs) produce viral particles and exhibit reverse transcriptase activity in human melanoma cells (Muster et al., 2003); TE expression is enhanced in urothelial and renal carcinoma cells (Florl et al., 1999), in human leukemia (Patzke et al., 2002; Depil et al., 2002), and in colorectal (Debniak et al., 2001) and human breast cancers (Wang-Johanning et al., 2003). Hypomethylation of LINEs and HERV-W retrotransposons, which is associated with high level of gene expression (Dante et al., 1992), has been observed in various cancers (Florl et al., 1999; Menendez et al., 2004; Ehrlich, 2003). It has been suggested that hypomethylation of TEs may promote genomic instability, and thus facilitate tumor progression (Xu and Deng, 2002).

Because TEs may be activated under certain conditions and may influence gene expression, they could be responsible for particular gene expression patterns in tumor tissues. Analyses of gene expression profiles in normal and tumor tissues have indeed revealed differential expression of particular genes (Schaner et al., 2003; Bucca et al., 2004; Taniwaki et al., 2006). For example, some genes are highly expressed in epithelial ovarian cancer, whereas under normal conditions they are generally almost silent (Welsh et al., 2001). These findings indicate that changes in gene expression occur during tumor development, but the direct cause of this deregulation has typically not been identified.

In this paper, we use the complete sequence of the human genome to detect TEs in the neighborhood of genes. We analyze the expression of human genes in 19 pairs of normal and tumor tissues. We check whether the presence of a TE is associated with a change in gene expression between normal and tumor tissues, which could indicate that TEs play a role in gene deregulation in cancer. We show that the presence of TEs is correlated to the average number of normal tissues in which a gene is expressed. We find that genes associated with SINE elements are particularly deregulated under tumor conditions, and that this deregulation increases with the number of SINE copies.

2. Materials and methods

2.1. Expression and genome data

We obtained six million ESTs from human tissues from GenBank (release October 2004 (Benson et al., 2004)). To allow us to assess the tissue origin of each EST with sufficient accuracy, we excluded cDNA libraries based on cell cultures, pooled organs, or unidentified tissues. After pooling all ESTs from libraries corresponding to the same tissue type, we retained only tissues that had been sampled with at least 10,000 ESTs to limit stochastic variations in expression measures. A total of 19 of these tissues were represented in both normal and tumor states and were thus included in our analysis. We downloaded transcribed sequences from Ensembl and retained one randomly chosen transcript per gene (version 24, October 2004, Hubbard et al., 2005). The association between genes and ESTs was determined by comparing CDS to the EST dataset using MEGABLAST (Zhang et al., 2000). We retained alignments showing at least 95% identity over 100 nucleotides