Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 365 (2006) 104–110
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Transposable elements as a significant source of transcription regulating signals

Bartley G. Thornburg [a,1], Valer Gotea [a,b,1], Wojciech Makałowski [a,b,c,*]

[a] Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
[b] Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received 6 May 2005; received in revised form 6 September 2005; accepted 27 September 2005
Available online 10 January 2006

Abstract

Transposable elements (TEs) are major components of eukaryotic genomes, contributing about 50% to the size of mammalian genomes. TEs serve as recombination hot spots and may acquire specific cellular functions, such as controlling protein translation and gene transcription. The latter is the subject of the analysis presented. We scanned TE sequences located in promoter regions of all annotated genes in the human genome for their content in potential transcription regulating signals. All investigated signals are likely to be over-represented in at least one TE class, which shows that TEs have an important potential to contribute to pre-transcriptional gene regulation, especially by moving transcriptional signals within the genome and thus potentially leading to new gene expression patterns. We also found that some TE classes are more likely than others to carry transcription regulating signals, which can explain why they have different retention rates in regions neighboring genes.

Keywords: Transposable elements; Gene regulation; Promoters; Transcription factor binding sites

1. Introduction

More than four years after the publication of the first draft of the human genome (Lander et al., 2001), scientists continue to face unsolved mysteries related to its structure. Among these, the abundance of transposable elements (TEs), which contributed to about half of the human genome (Makalowski, 2001; Lander et al., 2001), has no immediate rational explanation. There are many successful organisms with compact genomes, e.g. all prokaryotes, *Takifugu rubripes* among vertebrates, or *Arabidopsis thaliana* among flowering plants, and as a consequence, many scientists regarded these elements as "junk" (Ohno, 1972), unnecessary ballast, genomic burden, selfish DNA or parasites (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Hickey, 1982; Schmid, 2003). It was through the progress of the human genome project that knowledge about function of different genomic components increased significantly, including knowledge about origin and role of non-coding sequences (Hardison, 2000). More and more biologists started to regard repetitive elements as a genomic treasure (Brosius, 1991; Makalowski, 1995; Britten, 1996b; Brosius, 1999; Makalowski, 2003), as objects worthy of biological studies. Recent years witnessed accelerated progress in understanding genomic dynamics, and it appears that different mobile elements play a significant role in this process (Makalowski, 1995; Britten, 1996a; Lorenc and Makalowski, 2003; Brosius, 2005).

One of the most direct influences of transposable elements on the host genome is their role in modulating the structure and expression of "resident" genes. After discovery that long terminal repeats (integral parts of some retroelements) carry promoter and enhancer motifs it became clear that integration of such elements in proximity of a host gene must have an influence on this gene expression (Sverdlov, 1998). Many TEs have been described in the last decade that can add a variety of functions to their targeted genes. These include polyadenylation sites, promoters, enhancers, and silencers (Makalowski, 1995). It seems

Abbreviations: TE, transposable element; SINE, Short INterspersed Element; LINE, Long INterspersed Element; LTR, long terminal repeat; bp, base pair; pol, polymerase

* Corresponding author. Department of Biology, Pennsylvania State University, 514 Mueller Lab, University Park, PA 16802, USA.
E-mail address: wojtek@psu.edu (W. Makałowski).
[1] These authors contributed equally to the work.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.036

that a sizable fraction of eukaryotic, gene-associated regulatory elements arose in this modular fashion by insertion of TEs, and not only by point mutations of static neighboring sequences. When a TE is inserted upstream from a gene, a few short motifs can be conserved if they were subjected to selective pressure as promoters or enhancers of transcription. Even though the rest of the TE sequence might evolve beyond recognition due to absence of functional constraints, TEs are hence exapted into a novel function (Brosius and Gould, 1992). A recent survey that analyzed 846 functionally characterized *cis*-regulatory elements from 288 genes, showed that 21 of those elements (~2.5%) from 13 genes (~4.5%) reside in TE-derived sequences (Jordan et al., 2003). The same study showed that TE-derived sequences are present in many more (~24%) promoter regions, defined as ~500 bp located 5' of functionally characterized transcription initiation site. Similarly, van de Lagemaat et al. showed that the 5′ UTRs of a large proportion of mammalian mRNAs contain TE fragments, suggesting that they play a role in regulation of gene expression (van de Lagemaat et al., 2003). One should note that the TE influence on gene regulation upon insertion in promoter regions is only due to chance similarity of TE sequence to various *cis*-regulatory elements, or to the presence of regulatory elements that were active in regulating the transcription of the TE itself. To evaluate their content in such elements, we scanned TE sequences located in promoter regions of all annotated human genes for their content in putative transcription regulating signals. We found that not all regulatory signal classes are over-represented in TE-derived sequences as compared to randomly generated sequences of similar length and GC content, and that different TE classes greatly differ in their potential to fortuitously deliver regulatory signals upon insertion in gene promoter regions. Nevertheless, it is clear that all TEs have a potential to alter gene regulation given their mobility, with possible significant long term evolutionary consequences.

2. Materials and methods

2.1. Finding TEs in promoter sequences

For the purpose of this study we used the July 2003 assembly of the human genome available from the Golden Path at the University of California Santa Cruz (http://genome.ucsc.edu/goldenPath/hg16/), and corresponding gene annotation (we used the refFlat files which contain annotation for RefSeq and predicted genes). For every gene, we extracted 2000 nucleotides upstream from the annotated transcription start coordinate. The 20,193 excised promoter sequences were then scanned for occurrence of TEs using the May 15, 2002 version of RepeatMasker (http://www.repeatmasker.org) with default options, but ignoring simple repeats and low complexity regions ("-nolow" parameter).

2.2. Identification of transcription signals

TRANSFAC database of transcription factor binding sites, maintained by Biobase (http://www.biobase.de), was used as a source of verified transcription signals. We relied upon the MATCH program (Kel et al., 2003) from the same software suite for finding such putative signals in human promoter regions. MATCH uses predefined positional weight matrices (PWM), which we chose based on the TRANSFAC classification of transcription factor binding sites (http://www.gene-regulation.com/pub/databases/transfac/cl.html). Representative high-quality matrices were chosen for each class

Table 1
Representative position weight matrices (PWM) from TRANSFAC database used for identifying transcription factor binding sites in human promoter regions

Class	Factor name	Matrix ID	Quality	Matrix similarity cutoff[a]
Superclass: basic domains				
Leucine zippers	XBP-1	V$AP1_C	High	0.98
	CRE-BP1	V$CREBP1_Q2	High	0.96
	C/EBPα	V$CEBP_C	High	0.93
Helix–loop–helix	E12	V$E12_Q6	High	0.97
	MyoD	V$MYOD_01	High	0.94
Helix–loop–helix/ leucine zipper	USF	V$USF_Q6	High	0.95
	c-Myc	V$MYCMAX_01	High	0.97
RF-X	RF-X2	V$RF-X1_01	High	0.94
Helix–span–helix	AP-2γ	V$AP2_Q6_01	Low	0.92
Superclass: zinc-coordinating domains				
Zinc finger–nuclear receptor	GR	V$GRE_C	High	0.92
	ER	V$ER_Q6	High	0.94
	HNF-4α1	V$HNF4_01	High	0.86
Cys4 zinc fingers	GATA-1	V$GATA1_02	High	0.97
	GATA-3	V$GATA_C	High	0.96
Cys2His2 zinc fingers	YY1	V$YY1_02	High	0.92
	Egr-1	V$EGR1_01	High	0.96
Superclass: helix–turn–helix				
Homeo domain	HNF-1A	V$HNF1_01	High	0.90
	Oct-2B	V$OCT_C	High	0.93
Paired box	Pax-6	V$PAX6_01	High	0.88
	Pax-5	V$PAX_Q6	High	0.86
Fork head/winged helix	HNF3-α	V$HNF3B_01	High	0.94
	E2F-1	V$E2F_Q6	High	0.91
Tryptophan clusters	c-ETS-1 p54	V$ETS1_B	High	0.94
	IRF-1	V$IRF1_01	High	0.97
Superclass: beta-scaffold factors				
Rel homology region	p50	V $NFKAPPAB_01	High	0.96
	p65	V$NFKB_Q6_01	High	0.91
STAT	p91	V$STAT_01	High	0.97
MADS box	MEF-2A	V$MEF2_02	High	0.93
	SRF	V$SRF_C	High	0.93
TATA binding proteins	TBP	V$TATA_C	High	0.95
HMG	Sox-9	V$SOX9_B1	High	0.95
	SSRP1	V$TCF4_Q5	High	0.98
Heteromeric CCAAT factors	CP1B	V$NFY_Q6	High	0.96
Grainyhead	CP2	V$CP2_02	High	0.93
Runt	AML-3	V$AML_Q6	High	0.97

[a] The matrix similarity cutoff corresponds to a false negative rate of 50% (FN50).