POLYADENYLATION REGULATORY SEQUENCES AND FREQUENCY OF ALTERNATIVE POLYADENYLATION SITES IN A COMPREHENSIVE SET OF CANCER PREDISPOSITION GENES

VIEIRA, IGOR ARAUJO[1,2]; RECAMONDE-MENDOZA, MARIANA[3]; SILVA, VANDECLECIO LIRA DA[4,5]; LEÃO, DELVA PEREIRA[1,6]; SCHEID, MARINA ROBERTA[2]; SOUZA, SANDRO JOSÉ DE[4,5]; ASHTON-PROLLA, PATRICIA[1,2,6,7]

[1] Programa de Pós-graduação em Genética e Biologia Molecular (Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil); [2] Laboratório de Medicina Genômica (Serviço de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre - HCPA, Porto Alegre, Rio Grande do Sul, Brazil); [3] Instituto de Informática (UFRGS, Porto Alegre, Rio Grande do Sul, Brazil); [4] Programa de Pós-Graduação em Bioinformática (Universidade Federal do Rio Grande do Norte - UFRN, Natal, Rio Grande do Norte, Brazil); [5] Instituto do Cérebro (UFRN, Natal, Rio Grande do Norte, Brazil); [6] Serviço de Genética Médica (HCPA, Porto Alegre, Rio Grande do Sul, Brazil); [7] Departamento de Genética (UFRGS, Porto Alegre, Rio Grande do Sul, Brazil).

Almost all eukaryotic mRNAs acquire a poly(A) tail at their 3' ends in a process termed polyadenylation. Two core polydenylation elements (CPE) located in the 3' untranslated region (3'UTR) of pre-mRNAs play an essential role in this process: the polyadenylation signal (PAS), a highly conserved hexamer AAUAAA or its close variants; and the actual cleavage site (CS), preferentially a CA dinucleotide 10-30 nucleotides downstream of the PAS. In human genes, PAS sequences include 12 functional hexamer variants, because some positions are tolerant to point mutations. In addition, alternative polyadenylation (APA) is defined as use of more than one functional PAS/CS, allowing a single gene to encode multiple mRNA transcripts with variable 3'UTR. In the present study, we characterized PAS and CS sequences in a comprehensive set of cancer predisposition genes (CPGs), besides exploring the occurrence of APA events in the same group of genes. NCBI (reference source) and APA databases (APADB and APASdb) were queried to characterize CPE sequences in the selected CPGs (n=117), including 81 tumor suppressor genes and 17 oncogenes. Regarding CPGs with no PAS described in the NCBI database, we developed a computational method using in-house Perl scripts to identify the 3'-most hexamers that may function as PAS (putative PAS) in the full sequence of corresponding transcripts. Based on NCBI analysis, we did not find an established PAS in 21 of the 117 CPGs (~18%), and most PAS already described (74.4%) had the canonical sequence AAUAAA, while 24.1% contained the variant AUUAAA. Our computational strategy was able to detect putative PAS sequences for 17/21 CPGs with no established PAS in NCBI database, and putative PAS were not identified for the *ERCC4*, *FH*, *MUTYH* and *SHOC2* genes. Interestingly, we found the AA dinucleotide in most CS sequences associated with this set of CPGs. CA dinucleotide was only the fourth most frequent CS in our gene set, indicating that certain estimates provided by long-standing polyadenylation studies do not apply to all human transcripts. Moreover, an integrative analysis of the data obtained through the NCBI, APADB and APASdb databases allowed to identify 105 CPGs (~90%) with APA sites among their transcript variants, while a previous estimate indicated that it occurs in about 54% of human genes, suggesting a greater complexity in the regulation of polyadenylation in transcripts derived from CPGs. The strongest evidence of APA arose from the *PTEN* transcript which has 61 APA sites differentially used in its processing among different normal human tissues according to APASdb analysis. Overall, our study generated a landscape of polyadenylation regulatory sequences in CPGs that may be useful in the development of molecular analyses covering these often neglected regulatory elements of 3' end processing in human genes. These findings reinforce the relevance of establishing updated methods and/or databases to detect PAS and CS sequences. Furthermore, the computational strategies used here could be easily applied to similar situations with additional genes outside the CPG context. This is the first study focused on the comprehensive characterization of CPE sequences in CPGs.

Keywords: polyadenylation, cancer predisposition genes, alternative polyadenylation.

Acknowledgments: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).