A comprehensive analysis of core polyadenylation sequences and regulation by microRNAs in a set of cancer predisposition genes

Igor Araujo Vieira, Mariana Recamonde-Mendoza, Vandeclecio Lira da Silva, Delva Pereira Leão, Marina Roberta Scheid, Sandro José de Souza, Patricia Ashton-Prolla

PII: S0378-1119(19)30593-1
DOI: https://doi.org/10.1016/j.gene.2019.143943
Article Number: 143943
Reference: GENE 143943

To appear in: *Gene*

Received date: 27 January 2019
Revised date: 18 June 2019
Accepted date: 20 June 2019

Please cite this article as: I.A. Vieira, M. Recamonde-Mendoza, V.L. da Silva, et al., A comprehensive analysis of core polyadenylation sequences and regulation by microRNAs in a set of cancer predisposition genes, Gene, https://doi.org/10.1016/j.gene.2019.143943

A comprehensive analysis of core polyadenylation sequences and regulation by microRNAs in a set of cancer predisposition genes

Igor Araujo Vieira[a,b,*], Mariana Recamonde-Mendoza[c,d], Vandeclecio Lira da Silva[e], Delva Pereira Leão[a,d], Marina Roberta Scheid[b,f], Sandro José de Souza[e,g,h], Patricia Ashton-Prolla[a,b,f,i,j]

[a] Programa de Pós-graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[b] Laboratório de Medicina Genômica, Serviço de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil

[c] Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[d] Núcleo de Bioinformática, Serviço de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil

[e] Programa de Pós-Graduação em Bioinformática, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil

[f] Programa de Pós-graduação em Ciências Médicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

[g] Instituto do Cérebro, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil

[h] Bioinformatics Multidisciplinary Environment, Universidade Federal do Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil

[i] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil

[j] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

***Corresponding author:**
Igor Araujo Vieira
Laboratório de Medicina Genômica, Serviço de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre
Ramiro Barcelos, 2350
90035-903 Porto Alegre, Rio Grande do Sul, Brazil
Telephone: +55 51 3359.7661
Fax number: +55 51 3359.8761
e-mail: igvieira@hcpa.edu.br

Abstract

Two core polyadenylation elements (CPE) located in the 3' untranslated region of eukaryotic pre-mRNAs play an essential role in their processing: the polyadenylation signal (PAS) AAUAAA and the cleavage site (CS), preferentially a CA dinucleotide. Herein, we characterized PAS and CS sequences in a set of cancer predisposition genes (CPGs) and performed an *in silico* investigation of microRNAs (miRNAs) regulation to identify potential tumor-suppressive and oncogenic miRNAs. NCBI and alternative polyadenylation databases were queried to characterize CPE sequences in 117 CPGs, including 81 and 17 known tumor suppressor genes and oncogenes, respectively. miRNA-mediated regulation analysis was performed using predicted and validated data sources. Based on NCBI analyses, we did not find an established PAS in 21 CPGs, and verified that the majority of PAS already described (74.4%) had the canonical sequence AAUAAA. Interestingly, “AA” dinucleotide was the most common CS (37.5%) associated with this set of genes. Approximately 90% of CPGs exhibited evidence of alternative polyadenylation (more than one functional PAS). Finally, the mir-192 family was significantly overrepresented as regulator of tumor suppressor genes (P<0.01), which suggests a potential oncogenic function. Overall, this study provides a landscape of CPE in CPGs, which might be useful in development of future molecular analyses covering these frequently neglected regulatory sequences.

Keywords: polyadenylation; polyadenylation sequences; microRNAs; gene expression regulation; cancer predisposition genes.

1. Introduction

Polyadenylation, which comprises the pre-mRNA cleavage followed by adding a stretch of adenosine residues (called poly(A) tail) to the 3' end, is an essential nuclear step during pre-mRNAs processing in almost all eukaryotic cells [1, 2]. Three *cis*-acting RNA sequence elements often termed as the "core" polyadenylation elements (CPE) determine precisely the 3' end cleavage/polyadenylation site in mammalian pre-mRNAs: (1) the poly(A) signal (PAS), a highly conserved hexamer AAUAAA or its close variants; (2) the *de facto* polyadenylation site or cleavage site (CS), preferentially a "CA" dinucleotide located 10-30 nucleotides (nt) downstream of the PAS; and (3) a less well-conserved G-U-rich sequence 14-70 nt downstream of the PAS [3-7]. There are five major *trans*-acting protein factors that constitute the machinery involved in this process: cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), cleavage factors I and II (CFI and CFII), and poly(A) polymerase (PAP) [1, 2, 8]. The first one recognizes the AAUAAA sequence and CstF binds to the G-U-rich sequence, while CFI and CFII cleave RNA at the CS and generate a 3' end for PAP to perform polyadenylation [9-11]. Moreover, alternative polyadenylation (APA), defined as the use of more than one functional PAS/CS, occurs in at least 50% of human genes, allowing a single gene to encode multiple mRNA transcripts with variable 3' untranslated regions (3'UTR) [12, 13].

The known regulatory role of polyadenylation in mRNA localization, stability, and translation, as well as the emerging link between deleterious variants located in CPE and human disease underscore the need to fully characterize them [1, 14]. Alterations in CPE and neighboring sequences in the 3'UTR disrupt the cleavage and polyadenylation steps, resulting in several human pathologies [14-16]. Recently, functional CPE-related germline variants were reported in cancer predisposition genes (CPGs) and were associated with hereditary cancer syndromes (HCS). For example, the single nucleotide polymorphism (SNP) rs78378222 changes the constitutive PAS of the tumor suppressor gene *TP53* [17] and has been linked with a Li-Fraumeni-like syndrome phenotype [18]. Another functional variant (duplication of 20 base

pairs) located upstream of the *MSH6* PAS (a DNA-repair gene) was identified in two patients fulfilling clinical criteria for Lynch syndrome [19].

MicroRNAs (miRNAs) are endogenous, single-stranded, small non-coding RNAs (18-25 nt) that post-transcriptionally regulate gene expression through mRNA degradation and/or translation repression. In addition to CPE sequences, miRNA binding sites represent another class of elements predominantly located in the 3'UTR of mRNAs [20, 21]. Remarkable studies showed that miRNA mutations or altered expression levels correlate with various human cancers, and indicate that some miRNAs can function as tumor suppressors and/or oncogenes [22-25]. The deletion or downregulation of a miRNA that has a tumor suppressor role, by modulating the expression of oncogenes, leads to inappropriate amounts of the miRNA-target oncoproteins. The overall outcome might involve increased cellular proliferation, ultimately contributing to carcinogenesis. Additionally, oncogenic miRNA amplification or overexpression can reduce or eliminate the expression of a miRNA-target tumor-suppressor gene, resulting in an abnormal cellular proliferation and potentially, tumor formation [25].

In the present study, we aimed to characterize CPE (specifically PAS and CS) in a comprehensive set of CPGs and performed an *in silico* investigation of miRNA regulation in the same group of genes to identify miRNAs with potentially oncogenic or tumor suppressor function. These analyses generated a CPE landscape in CPGs that can be used in the development of more comprehensive diagnostic methods, including these regulatory sites often neglected during the routine molecular analysis. Furthermore, computationally predicted and experimentally validated data allowed us to suggest a potential oncogenic function of the mir-192 family.

2. Results

2.1. Characterization of core polyadenylation sequences in cancer predisposition genes

The number and sequence of constitutive PAS and CS in the longest variant transcript derived from each CPG are shown in **Table 1**. Distance

(measured in base pairs) between these specific CPE and analysis of the presence of APA sites are also indicated. Importantly, a PAS sequence was not found in 21 of the 117 CPGs studied (~18%) using National Center for Biotechnology Information (NCBI) database (reference source). Among these 21 genes, several were associated with highly prevalent HCS or related phenotypes, namely *BRCA2* (Hereditary Breast and Ovarian Cancer Syndrome), *BRIP1* (Familial Breast and Ovarian Cancer), *CDK4* and *TERT* (Familial Malignant Melanoma), *MEN1 (*Multiple Endocrine Neoplasia Type 1), and *PMS2 (*Lynch Syndrome).

Table 1. Characterization of polyadenylation regulatory sequences in cancer predisposition genes included in this study.

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
APC (TSG)	Familial Adenomatous Polyposis, including attenuated phenotype; Gardner Syndrome and Turcot's Syndrome (mainly colorectal cancer)	1 (CATAAA)	1 (CA)	16	6	18
ATM (TSG)	Ataxia-telangiectasia (breast cancer, leukemia and lymphoma)	1 (AATAAA)	1 (GC)	12	4	31
BAP1 (TSG)	Predisposition to malignant mesothelioma, cutaneous and uveal melanoma, and renal cell carcinoma	1 (AATAAA)	1 (AA)	28	2	6
BARD1 (TSG)	Hereditary Breast and Ovarian Cancer Syndrome	1 (AATAAA)	1 (TG)	16	2	7
BLM (TSG)	Bloom Syndrome (leukemia, lymphoma and colorectal cancer)	1 (AATAAA)	2 (Both AA)	13-16	3	7
BMPR1A (TSG)	Juvenile Polyposis Syndrome (colorectal cancer and other tumors of the gastrointestinal tract)	1 (AATAAA)	1 (GA)	18	2	10
BRAF (ONC)	LEOPARD, Cardiofaciocutaneous and Noonan syndromes (predisposition to various tumor types); predisposition to melanoma	3 (All hexamers are AATAAA)	3 (All CS are AA)	18-26	2	18
BRCA1 (TSG)	Hereditary Breast and Ovarian Cancer Syndrome	1 (AATAAA)	1 (CA)	18	1	5
BRCA2 (TSG)	Hereditary Breast and Ovarian Cancer Syndrome and Fanconi Anemia (predisposition to leukemia and certain solid tumors)	NI	NI	NI	1	11
BRIP1 (TSG)	Fanconi Anemia; familial breast and ovarian cancer	NI	1 (AA)	NI	NI	7
CDC73 (TSG)	Familial Parathyroid Cancer (often called Hyperparathyroidism-Jaw Tumor Syndrome)	4 (2 ATTAAA and 2 AATAAA)	4 (2 AA, TA and TC)	8-26	10	31
CDH1 (TSG)	Hereditary Gastric Cancer (predisposition to gastric and colorectal tumors)	1 (ATTAAA)	1 (AT)	16	1	10

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
CDKN1C (TSG)	Beckwith-Wiedemann Syndrome (predisposition to embryonal tumors and childhood tumors)	1 (AATAAA)	1 (GC)	17	NI	3
CDKN2A (TSG)	Melanoma-Pancreatic Cancer Syndrome	2 (Both AATAAA)	1 (CA)	21-27	2	4
CDK4 (ONC)	Familial Malignant Melanoma	NI	2 (AA and GC)	NI	1	5
CHEK2 (TSG)	Hereditary Breast and Colorectal Cancer	1 (ATTAAA)	1 (GG)	16	NI	2
DDB2 (TSG)	Xeroderma pigmentosum (predisposition to skin cancer, melanoma, leukemia and neoplasms in internal organs)	1 (AATAAA)	1 (AG)	14	1	9
ELAC2 (TSG)	Hereditary prostate cancer	2 (Both AATAAA)	2 (GA and TA)	17-19	3	7
EPCAM (UND)	Lynch Syndrome (often called Hereditary Nonpolyposis Colorectal Cancer)	2 (ATTAAA and AATAAA)	2 (GA and AA)	22-30	2	7
EPHB2 (TSG)	Predisposition to prostate cancer	2 (Both AATAAA)	2 (AA and TA)	13-17	2	7
ERCC2 (TSG)	Xeroderma pigmentosum	NI (transcript variant 1) and AATAAA (transcript variant 2)	GA (transcript variant 1) and AA (transcript variant 2)	42 (transcript variant 2)	2	2
ERCC3 (TSG)	Xeroderma pigmentosum	1 (AATAAA)	1 (GT)	21	1	3
ERCC4 (TSG)	Xeroderma pigmentosum	NI	NI	NI	3	17
ERCC5 (TSG)	Xeroderma pigmentosum	1 (AATAAA)	2 (TC and TA)	18-24	1	10
EXT1 (TSG)	Hereditary Multiple Osteochondromas and Bone Cancer Predisposition Syndrome (predisposition to chondrosarcoma)	2 (Both ATTAAA)	2 (GA and AA)	17-25	2	19
EXT2 (TSG)	Hereditary Multiple Osteochondromas and Bone Cancer Predisposition Syndrome	2 (Both ATTAAA)	2 (TT and CA)	18-22	2	16
FANCA (TSG)	Fanconi Anemia (predisposition to leukemia and certain solid tumors)	NI	NI	NI	4	5
FANCB (TSG)	Fanconi Anemia	NI	NI	NI	NI	NI
FANCC (TSG)	Fanconi Anemia	2 (ATTAAA and AATAAA)	2 (Both AA)	13-17	1	11
FANCD2 (TSG)	Fanconi Anemia	1 (AATAAA)	1 (TC)	14	2	5
FANCE (TSG)	Fanconi Anemia	1 (AATAAA)	1 (TG)	15	1	1

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
FANCF (TSG)	Fanconi Anemia	1 (AATAAA)	1 (CA)	16	6	6
FANCG (TSG)	Fanconi Anemia	1 (AGTAAA)	1 (AT)	12	1	1
FANCI (TSG)	Fanconi Anemia	NI	NI	NI	2	11
FANCL (TSG)	Fanconi Anemia	1 (AATAAA)	1 (AA)	16	1	6
FAS (UND)	Autoimmune Lymphoproliferative Syndrome (predisposition to lymphoma)	2 (Both ATTAAA)	2 (Both TA)	13-18	3	5
FASLG (UND)	Autoimmune Lymphoproliferative Syndrome	1 (AATAAA)	1 (AA)	20	4	1
FH (TSG)	Hereditary Leiomyomatosis and Renal Cell Carcinoma Syndrome	NI	NI	NI	1	5
FLCN (TSG)	Birt-Hogg-Dube Syndrome (predisposition to cutaneous hamartomas and renal tumors)	1 (AATAAA)	1 (AA)	15	3	3
GPC3 (TSG)	Simpson-Golabi-Behmel Syndrome (predisposition to embryonal tumors)	1 (AATAAA)	1 (AA)	14	NI	NI
GREM1 (ONC)	Juvenile Polyposis Syndrome and Hereditary Mixed Polyposis Syndrome (colorectal cancer)	1 (AATAAA)	1 (TG)	19	NI	5
HRAS (ONC)	Costello and Noonan Syndromes (predisposition to various tumor types)	NI	1 (GA)	NI	1	1
IL2RG (UND)	X-Linked Severe Combined Immunodeficiency (predisposition to lymphoma)	1 (AATAAA)	1 (AA)	23	1	NI
KIT (ONC)	Familial Gastrointestinal Stromal Tumors	1 (AATAAA)	1 (AA)	18	1	3
KRAS (ONC)	Cardiofaciocutaneous, Costello and Noonan Syndromes (predisposition to various tumor types)	NI	NI	NI	6	17
LIG4 (TSG)	DNA Ligase IV Syndrome (predisposition to leukemia and lymphoma)	1 (AATAAA)	1 (AA)	20	1	8
MAP2K1 (ONC)	Cardiofaciocutaneous and Noonan Syndromes	1 (AATAAA)	1 (AT)	12	1	4
MAP2K2 (ONC)	Cardiofaciocutaneous and Noonan Syndromes	NI	1 (AA)	NI	3	3
MAX (TSG)	Hereditary Pheochromocytoma and Paraganglioma	1 (AATAAA)	2 (Both TT)	18[e]	12	18

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
MC1R (UND)	Familial Malignant Melanoma	1 (AATAAA)	1 (TA)	23	1	3
MDM2 (ONC)	Related to Li-Fraumeni Syndrome (genetic modifiers; further information in *TP53* gene)	1 (AATAAA)	1 (CA)	14	7	26
MDM4 (ONC)		3 (2 ATTAAA and AATAAA)	3 (2 CA and TA)	16-22	7	30
MEN1 (TSG)	Multiple Endocrine Neoplasia Type 1	NI	1 (TA)	NI	NI	1
MET (ONC)	Hereditary Papillary Renal Cell Carcinoma	1 (ATTAAA)	1 (TA)	22	1	1
MITF (ONC)	Predisposition to malignant melanoma and renal cell carcinoma	1 (ATTAAA)	2 (CA and TA)	14-18	1	18
MLH1 (TSG)	Lynch Syndrome	1 (AATAAA)	1 (AA)	13	2	12
MRE11A (TSG)	Ataxia-telangiectasia-like disorder	2 (ATTAAA and AATAAA); overlapping PAS	4 (2 TA, CA and AA)	16-34	3	11
MSH2 (TSG)	Lynch Syndrome	1 (AATAAA)	1 (GA)	22	1	15
MSH6 (TSG)	Lynch Syndrome	1 (AATAAA)	1 (GA)	18	1	8
MSR1 (TSG)	Predisposition to prostate cancer	1 (ATTAAA)	1 (GA)	17	2	16
MUTYH (TSG)	Familial Adenomatous Polyposis 2 (Polyposis associated with *MUTYH*)	NI	1 (TA)	NI	1	2
NBN (TSG)	Nijmegen Breakage Syndrome (predisposition to leukemia, lymphoma and certain solid tumors)	2 (ATTAAA and AATAAA)	3 (All CS are AA)	17-36	3	17
NF1 (TSG)	Neurofibromatosis type 1 (predisposition to neurofibrosarcomas, optic gliomas, meningiomas and other malignant tumors)	1 (AATAAA)	2 (AA and GA)	23-1806	9	32
NF2 (TSG)	Neurofibromatosis type 2 (predisposition to vestibular schwannomas, meningiomas and other malignant tumors)	1 (AATAAA)	1 (TA)	18	2	8
NSD1 (UND)	Beckwith-Wiedemann Syndrome	1 (AATAAA)	3 (TA, AA, GA)	13-4469	3	23
PALB2 (TSG)	Fanconi Anemia; predisposition to breast and pancreas cancer	1 (AATAAA)	1 (TC)	18	1	2
PALLD (ONC)	Familial Pancreatic Adenocarcinoma	2 (Both AATAAA)	2 (AA and TA)	11-13	1	46

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
PMS1 (TSG)	Lynch Syndrome	2 (AATAAA and ATTAAA)	2 (CA and GA)	13-16	3	17
PMS2 (TSG)	Lynch Syndrome	NI	1 (GA)	NI	1	3
POLB (UND)	Alternative DNA polymerase; overexpression in certain solid tumors	1 (AATAAA)	1 (GA)	20	1	3
POLD1 (UND)	Alternative DNA polymerase; Polymerase Proofreading-associated polyposis; predisposition to colorectal, breast and endometrial cancer	1 (AATAAA)	1 (TG)	20	1	3
POLE (UND)	Alternative DNA polymerase; Polymerase Proofreading-associated polyposis; predisposition to colorectal, pancreas, ovarian and small intestine cancer	NI	1 (TC)	NI	1	4
POLH (UND)	Translesion DNA Polymerase; Xeroderma pigmentosum Variant Type (predisposition to skin cancer)	2 (Both AATAAA)	2 (CA and TA)	19-20	4	1
POLK (UND)	Translesion DNA Polymerase; overexpression in rectum and lung tumors	NI	NI	NI	2	16
POLQ (UND)	Translesion DNA Polymerase; overexpression in certain solid tumors	1 (AATAAA)	1 (TA)	20	NI	1
POT1 (TSG)	Familial Glioma and Melanoma	2 (Both AATAAA)	2 (AA and GA)	16-17	3	13
PRCC (UND)	Papillary Renal Cancer Syndrome	1 (AATAAA)	1 (CA)	18	2	3
PRKAR1A (TSG)	Carney Complex Disorder (predisposition to myxoid subcutaneous, testicular sertoli cell and thyroid cancer, and other solid tumors)	4 (3 AATAAA and 1 ATTAAA)	4 (TT, GA and 2 AA)	16-31	21	25
PRSS1 (UND)	Familial pancreatic cancer	1 (AATAAA)	1 (TC)	19	NI	1
PTCH1 (TSG)	Gorlin Syndrome, often called Basal Cell Nevus Syndrome (predisposition to basal cell carcinoma and medulloblastoma)	NI	NI	NI	1	10
PTEN (TSG)	Cowden Syndrome (predisposition to breast, thyroid, endometrial and other cancers)	5 (3 ATTAAA and 2 AATAAA)	6 (2 AA, 2 TA, GA and CA)	14-19	21	61
PTPN11 (UND)	Noonan and LEOPARD Syndromes	3 (2 AATAAA and 1 ATTAAA)	3 (2 GA and AA)	16-32	3	14

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
RAD50 (TSG)	Nijmegen Breakage-like Syndrome	1 (AATAAA)	2 (AA and CA)	16-741	7	31
RAD51 (TSG)	Fanconi Anemia and Familial Breast and Ovarian Cancer	2 (Both AATAAA)	4 (3 AA and GT)	20-80	NI	6
RAD51C (TSG)	Hereditary Breast and Ovarian Cancer	1 (AATAAA)	2 (CA and TA)	11-30	2	6
RAD51D (TSG)	Hereditary Breast and Ovarian Cancer	1 (AATAAA)	1 (AA)	15	2	4
RAF1 (ONC)	Noonan and LEOPARD 2 Syndromes	2 (Both AATAAA); overlapping PAS	1 (AA)	15-20	3	7
RB1 (TSG)	Hereditary Retinoblastoma	1 (AATAAA)	1 (TC)	15	6	21
RECQL4 (TSG)	Rothmund-Thomson Syndrome (predisposition to basal cell carcinoma, squamous cell carcinoma and osteosarcoma)	1 (AATAAA)	1 (GG)	23	1	1
RET (ONC)	Multiple Endocrine Neoplasia Type 2	1 (AATAAA or ATTAAA; isoform-specific PAS)	1 (TT or AA, isoform-specific CS)	11-15	NI	5
RNASEL (TSG)	Hereditary prostate cancer	1 (AATAAA)	1 (TA)	15	3	NI
SBDS (UND)	Shwachman-Diamond Syndrome (predisposition to myelodysplasia and acute myeloid leukemia)	1 (AATAAA)	1 (TT)	17	NI	1
SDHA (TSG)	Hereditary Paraganglioma-Pheochromocytoma	1 (AATAAA)	2 (AA and CA)	32-415	1	7
SDHAF2 (TSG)	Hereditary Paraganglioma-Pheochromocytoma	1 (ATTAAA)	3 (2 AA and TA)	18-42	2	5
SDHB (TSG)	Hereditary Paraganglioma-Pheochromocytoma	2 (Both AATAAA)	2 (Both AA)	15-29	1	1
SDHC (TSG)	Hereditary Paraganglioma-Pheochromocytoma	1 (ATTAAA)	2 (GA and AA)	17-1555	1	5
SDHD (TSG)	Hereditary Paraganglioma-Pheochromocytoma	2 (Both AATAAA)	2 (AA and TA)	0-18	1	1
SH2D1A (UND)	X-Linked Lymphoproliferative Disease	1 (AATAAA)	1 (AA)	24	2	NI
SHOC2 (ONC)	Cardiofaciocutaneous and Noonan Syndromes	NI	2 (TA and GA)	NI	3	19
SLX4 (TSG)	Fanconi Anemia	1 (AATAAA)	1 (AA)	19	1	2
SMAD4 (TSG)	Juvenile Polyposis Syndrome	3 (2 AATAAA and ATTAAA)	3 (2 AA and TT)	11-25	6	17
SMARCA4 (TSG)	Rhabdoid Tumor Predisposition Syndrome	2 (AATAAA and ATTAAA); overlapping PAS	2 (GA and CC)	20-25	4	10

Gene (ONC, TSG or UND)	Syndrome/Tumor predisposition	NCBI database (Constitutive PAS and CS sequences)[a]			Alternative polyadenylation (Number of poly(A) sites)	
		Number of PAS and sequence	Number of CS and sequence	Distance between the PAS and CS (bp)[b]	APADB[c]	APASdb[d]
SMARCB1 (TSG)	Rhabdoid Tumor Predisposition Syndrome	1 (AATAAA)	1 (CA)	30	1	1
SOS1 (ONC)	Noonan Syndrome	1 (AATAAA)	1 (AA)	24	5	47
STK11 (TSG)	Peutz-Jeghers Syndrome (tumors of the gastrointestinal tract, breast and ovarian cancer)	1 (AATAAA)	1 (TG)	19	3	4
TERT (UND)	Familial Malignant Melanoma	NI	NI	NI	NI	1
TMEM127 (TSG)	Familial Pheochromocytoma and predisposition to renal cell carcinoma	NI	1 (AA)	NI	9	10
TP53 (TSG)	Li-Fraumeni Syndrome (sarcoma, breast, brain and adrenocortical tumors, and other cancers)	1 (AATAAA)	1 (CA)	11	3	6
TSC1 (TSG)	Tuberous sclerosis (facial angiofibroma, renal angiomyolipoma, renal cell carcinoma, cardiac rhabdomyoma, and other tumors)	1 (AATAAA)	1 (AC)	19	4	19
TSC2 (TSG)	Tuberous sclerosis	2 (Both AATAAA); overlapping PAS	4 (2 GA and 2 CA, isoform-specific CS)	20-43	1	4
VHL (TSG)	Von Hippel-Lindau Syndrome (retina and central nervous system hemangioblastomas, and renal cell carcinoma) and Familial Pheochromocytoma	1 (ATTAAA)	1 (AA)	18	6	8
WAS (UND)	Wiskott-Aldrich Syndrome (predisposition to hematopoietic malignancies)	1 (AATAAA)	1 (AA)	26	3	NI
WRN (TSG)	Werner Syndrome (soft tissue sarcomas, osteosarcoma, meningioma, and other tumors)	2 (ATTAAA and AATAAA)	4 (2 AA and 2 TA)	16-24	2	13
WT1 (TSG)	Hereditary Wilms tumor (childhood renal cancer)	1 (ATTAAA)	1 (AA)	16	NI	6
XPA (TSG)	Xeroderma pigmentosum	1 (AATAAA)	1 (GA)	20	1	10
XPC (TSG)	Xeroderma pigmentosum	1 (AATAAA)	1 (AA)	12	3	12
XRCC3 (TSG)	Familial Malignant Melanoma	1 (AATAAA)	1 (CA)	14	1	2

ONC, oncogene; TSG, tumor suppressor gene; UND, "undetermined function" gene; PAS, polyadenylation signals; CS, cleavage sites; bp, number of base pairs; APADB, Alternative Polyadenylation Database; APASdb, Alternative Polyadenylation Sites Database; NI, not identified. [a] Constitutive PAS and CS indicated in NCBI database for the longest variant transcript (in bp) derived from each cancer predisposition gene; [b] Distance (in bp) between the PAS and CS calculated according their positions indicated in the NCBI interface; [c] Analysis using this database included only cleavage/polyadenylation sites identified from human peripheral blood samples (Müller *et al*., 2014, reference 61); [d] Analysis considering alternative CS mapped in all transcript variants (for each gene) from samples of 22 normal human tissues (You *et al*., 2015, reference 62); [e] One of the polyadenylation sites in the longest variant transcript derived from *MAX* gene (NM_145113.2) does not have a close PAS (upstream region).

Next, the frequency of different PAS hexamers was determined and compared to data from two previous genomic-scale studies [12, 26] (**Table 2**). About 74% of the described PAS for this set of CPGs exhibited the canonical hexamer AAUAAA, while approximately 24% had the functional variant PAS AUUAAA. These frequencies did not differ significantly between the two main groups of genes (oncogenes and tumor suppressor genes) (P=0.78). Moreover, we identified two CPGs with less frequent hexanucleotides representing their constitutive PAS: *APC* (CAUAAA) and *FANCG* (AGUAAA) (**Table 1**). Curiously, overlapping PAS were described for the *MRE11A*, *RAF1*, *SMARCA4*, and *TSC2* genes, and the distance calculated between PAS and CS sequences for thirteen CPGs (~11%) was greater than the ones established by long-standing experimental studies (10-30 bp) [3-5], suggesting that this theoretical convention does not apply to all human genes (**Table 1**).

Table 2. Comparison between the frequencies of human polyadenylation signal hexamers identified by previous genomic studies and in cancer predisposition genes analyzed in the current study.

Hexamer	Frequency, % (ranking)				
	Genomic data; reported by Tian *et al*., 2005[a]	Genomic data; reported by Beaudoing *et al*., 2000[b]	Cancer predisposition genes; N=96[c]	Oncogenes; N=12[d]	Tumor suppressor genes; N=69[e]
AAUAAA	53.18 (1)	58.20 (1)	74.45 (1)	73.68 (1)[f]	76.81 (1)[f]
AUUAAA	16.78 (2)	14.90 (2)	24.09 (2)	26.32 (2)[f]	20.29 (2)[f]
UAUAAA	4.37 (3)	3.20 (3)	NI	NI	NI
AGUAAA	3.72 (4)	2.70 (4)	0.73 (3)	NI	1.45 (3)
AAGAAA	2.99 (5)	1.10 (10)	NI	NI	NI
AAUAUA	2.13 (6)	1.70 (5)	NI	NI	NI
AAUACA	2.03 (7)	1.20 (8)	NI	NI	NI
CAUAAA	1.92 (8)	1.30 (6)	0.73 (3)	NI	1.45 (3)
GAUAAA	1.75 (9)	1.30 (7)	NI	NI	NI
AAUGAA	1.56 (10)	0.80 (11)	NI	NI	NI
UUUAAA	1.20 (11)	1.20 (9)	NI	NI	NI
ACUAAA	0.93 (12)	0.60 (13)	NI	NI	NI
AAUAGA	0.60 (13)	0.70 (12)	NI	NI	NI

NI, not identified.

[a] Frequency of hexamer/hexanucleotide types identified as polyadenylation signals (PAS) in the human genome as previously reported by Tian *et al*., 2005 [12]. The method developed by Beaudoing *et al*., 2000 [26] was applied to the human genome sequences located 1 to 40 nucleotides upstream of polyadenylation sites to detect hexamers that may function as PAS.

[b] Frequency of hexamer types identified as PAS in the human genome using a method described by Beaudoing *et al*., 2000 [26].

[c] Frequency considering PAS indicated in NCBI for 96 of 117 genes included in this study (the remaining genes have no description of PAS in this database). Each hexamer was counted as a unit, including the case of transcripts with more than one PAS sequence and transcripts derived from *RET* gene that have different hexamers (AAUAAA and AUUAAA) in their specific isoforms.

[d] Frequency considering PAS indicated in NCBI for 12 of 17 oncogenes included in this study. Two putative oncogenes were included in this group (*PALLD* and *SHOC2*).

[e] Frequency considering PAS indicated in NCBI for 69 of 81 tumor suppressor genes (TSG) included in this study. Eight putative TSG were included in this group (see Materials and Methods).

[f] Frequency comparison of most common hexamers functioning as PAS (AAUAAA and AUUAAA) between oncogenes and tumor suppressor genes: P=0.78 (Fisher's exact test).

An integrative analysis of data retrieved from NCBI, APADB, and APASdb databases allowed us to identify 105 CPGs (~90%) with evidence of APA among their transcript variants, considering that more than one functional PAS/CS had been indicated by at least one of these databases (based on the results presented in **Table 1**). However, it is important to emphasize that some CPGs included in this study had little or no previous data regarding their CPE sequences (e.g., *FANCB*, *HRAS*, *MEN1*, and *TERT*), rendering the evaluation of APA site occurrence impossible in these genes. 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 (**Fig. 1**). Interestingly, although experimental evidence suggests a large number of APA sites in the *PTEN* gene, two CS located in distinct exonic regions (chr10:89726131 and chr10:89726870) showed higher usage quantification (29.4% and 12.3%, respectively). Both are associated with PAS containing the functional AUUAAA hexamer (**Fig. 1**) and are expected to be preferentially used by the cleavage/polyadenylation protein complex.

Regarding the 21 CPGs without PAS sequences described in the NCBI database, our computational strategy was able to detect 3'-most hexamers that might function as a PAS for 17 of these genes (**Table 3**). Although the same method was employed to screen the full sequence of corresponding transcripts, we were unable to identify putative PAS for *ERCC4*, *FH*, *MUTYH*, and *SHOC2*.

Table 3. Putative polyadenylation signals for cancer predisposition genes without identification of this sequence in NCBI database.

Gene	Chr number	RefSeq	Transcript start	Transcript end	PAS sequence[a]	PAS start	PAS end
BRCA2	chr13	NM_000059	32889616	32973809	AATAAA	32973352	32973357
BRCA2	chr13	NM_000059	32889616	32973809	ATTAAA	32973676	32973681
BRIP1	chr17	NM_032043	59756546	59940920	AATAAA	59938913	59938918
BRIP1	chr17	NM_032043	59756546	59940920	ATTAAA	59939856	59939861
CDK4	chr12	NM_000075	58141509	58146230	AGTAAA	58145246	58145251
CDK4	chr12	NM_000075	58141509	58146230	CATAAA	58144382	58144387
ERCC2	chr19	NM_000400	45854648	45873845	ATTAAA	45871095	45871100
ERCC2	chr19	NM_000400	45854648	45873845	AATACA	45871113	45871118
FANCA	chr16	NM_000135	89803958	89883065	AATAAA	89878113	89878118
FANCA	chr16	NM_000135	89803958	89883065	ATTAAA	89882343	89882348
FANCB	chrX	NM_001018113	14861528	14891191	AATAAA	14889174	14889179
FANCB	chrX	NM_001018113	14861528	14891191	TATAAA	14890834	14890839
FANCI	chr15	NM_001113378	89787193	89860362	AATAAA	89860342	89860347
FANCI	chr15	NM_001113378	89787193	89860362	TATAAA	89857626	89857631
HRAS	chr11	NM_001130442	532241	535567	AGTAAA	532259	532264
HRAS	chr11	NM_001130442	532241	535567	AATATA	534310	534315
KRAS	chr12	NM_004985	25357722	25403865	AATAAA	25401738	25401743
KRAS	chr12	NM_004985	25357722	25403865	TATAAA	25403182	25403187
MAP2K2	chr19	NM_030662	4090319	4124126	AATAAA	4118828	4118833
MAP2K2	chr19	NM_030662	4090319	4124126	ATTAAA	4114316	4114321
MEN1	chr11	NM_130800	64570985	64578035	TATAAA	64572816	64572821
MEN1	chr11	NM_130804	64570985	64578766	AGTAAA	64575874	64575879
PMS2	chr7	NM_000535	6010555	6048737	AATAAA	6047840	6047845
PMS2	chr7	NM_000535	6010555	6048737	ATTAAA	6047310	6047315
POLE	chr12	NM_006231	133200347	133264110	AATAAA	133257497	133257502
POLE	chr12	NM_006231	133200347	133264110	TATAAA	133256791	133256796
POLK	chr5	NM_016218	74807656	74895646	AATAAA	74895618	74895623
POLK	chr5	NM_016218	74807656	74895646	ATTAAA	74894636	74894641
PTCH1	chr9	NM_001083607	98205263	98269481	ATTAAA	98268733	98268738
PTCH1	chr9	NM_001083607	98205263	98269481	TATAAA	98265099	98265104
TERT	chr5	NM_001193376	1253286	1295162	AATAAA	1269949	1269954
TERT	chr5	NM_001193376	1253286	1295162	AGTAAA	1274131	1274136
TMEM127	chr2	NM_001193304	96915945	96931751	ATTAAA	96929713	96929718
TMEM127	chr2	NM_001193304	96915945	96931751	TATAAA	96929395	96929400

Chr Number, chromosome number; PAS, polyadenylation signal.

[a] A computational method was used to screen 3'-most canonical AAUAAA hexamers and its human functional variants in the full sequence of corresponding transcripts (RefSeq sequences).

In addition to the investigation of PAS hexamer distribution, we also assessed the frequency of different dinucleotide sequences identified as cleavage sites (CS) in CPGs (**Table 4**). It is well known that most pre-mRNAs in eukaryotes are preferentially cleaved immediately downstream from a "CA" dinucleotide [5, 27]. However, "AA" was the most common sequence (37.5%) used by this specific set of genes. Surprisingly, the "CA" dinucleotide was only the fourth most frequent CS (13.06%) in our analysis.

Table 4. Frequency of dinucleotide sequences identified as cleavage sites in cancer predisposition genes analyzed in the current study.

Dinucleotide sequence	Frequency in transcripts analyzed in this study (%)[a]
AA	37.50
TA	17.05
GA	14.20
CA	13.06
TC	3.98
TT	3.98
TG	2.85
AT	1.71
GC	1.71
GG	1.14
GT	1.14
AC	0.56
AG	0.56
CC	0.56

[a] Frequency calculated considering the cleavage sites (CS) indicated in NCBI for 107 out of 117 genes included in this study (the remaining genes have no description of CS in this database). Each dinucleotide was counted as a unit, including the case of genes with more than one CS sequence and transcripts derived from *ERCC2*, *RET*, and *TSC2* genes that have different CS sequences in their specific isoforms.

2.2. *In silico* identification of potentially tumor-suppressive and oncogenic miRNAs

Based on evidences from experimental validation and computational prediction, we investigated miRNAs/miRNA families that modulate the expression of a subset of CPGs included in this study: 81 tumor suppressor genes and 17 oncogenes. First, 1396 and 184 miRNAs from experimentally validated and predicted interactions, respectively, were identified as regulators of both gene groups, and the number of intersections among these sets of miRNAs was defined (**Fig. S1**). Considering the filtering criteria employed in these *in silico* analyses, only the *FANCB* gene had no miRNA regulation reported in our data. We then performed an over-representation analysis to identify potentially oncogenic or tumor suppressor miRNAs. **Fig. 2** shows the regulatory networks obtained by combining analysis of both validated and predicted data, indicating miRNAs and miRNA families significantly overrepresented (P<0.01) as regulators of tumor suppressor genes and oncogenes. As seen in these interactomes, statistically significant data were derived mainly from experimental sources (**Figs. 2A and 2B**). Interestingly, experimental data also represented the majority of collected miRNA-target gene interactions compared to the data only predicted by computational tools (**Figs. 2C and 2D**).

Lastly, we observed that the mir-192 family was the most significantly overrepresented among tumor suppressor genes (P=0.002), which could suggest an oncogenic function. This miRNA family has experimentally validated binding sites in the 3'UTR of 20 tumor suppressor genes included in this study (**Fig. 2B**). Among them, there are central genes in human DNA repair pathways such as *BRCA1*, *BRCA2*, *MSH6*, *RAD51*, and *XPA*. Although less oncogenes were included, a larger number of miRNA families were strongly associated with these genes: mir-128, mir-1471, mir-483, mir-3170 and mir-218. An additional result, important to validate our analysis, refers to mir-34a-5p, a miRNA already known to have tumor suppressor function [28, 29]. This miRNA family was significantly overrepresented as regulator of oncogenes (P=0.005), and in our analyses we observed validated and predicted interactions with 8 of 17 oncogenes studied (**Fig. 2A**).

3. Discussion

Recent advances have shown that eukaryotic cleavage and polyadenylation mechanisms are regulated through a network of *cis*-acting RNA sequence elements (herein termed CPE) located at the pre-mRNA 3'UTR and *trans*-acting proteins, contributing to the qualitative/quantitative adjustment of gene expression. The CPE arrangement determines the efficiency of a given polyadenylation site [2, 16]. The most prominent CPE is the PAS, a hexameric sequence motif located 10–30 nt upstream of the CS that was first described by Proudfoot and Brownlee (1976) [3]. The PAS serves as a binding site for CPSF, an endonuclease responsible for pre-mRNA cleavage [8]. Previous studies indicated that in around 55% of human mRNAs, the PAS hexamer is the AAUAAA consensus sequence [12, 26], recognized as one of the most highly conserved sequence elements known [3, 11, 30, 31]. Most of the remaining mRNAs (~45%) that do not have an exact match to the consensus differ by only a single substitution. An A → U conversion at the second position is the most common PAS variant (AUUAAA) [12, 32]. Recently, apart from the 12 hexamer variants previously identified (mentioned in the Table 2) [12, 26], six novel motifs conserved between human and mouse were suggested as potential PAS sequences [33].

In this regard, recent studies identified functional germline variants located within or in the vicinity of CPE sequences in two CPGs [17,19], raising the necessity to characterize these elements. In the current study, we evaluated the frequency and sequence of PAS hexamers and CS dinucleotides in a set of CPGs, as well as the distance (in bp) between them. In agreement with some previous genomic scale data [12, 26], our analysis revealed that the majority of PAS contained the canonical hexamer AAUAAA, and the AUUAAA variant was the second most frequent. However, we found the "AA" dinucleotide in most CS sequences associated with CPGs, which has not been previously reported. Although the nucleotide sequence of the exact CS is not highly conserved [34], most pre-mRNAs are cleaved downstream of an adenosine residue (in agreement with our data) and "CA" was defined as the optimal CS [5, 35]. Interestingly, "CA" dinucleotide was only the fourth most frequent CS in our gene set. Since more than 50% of human protein coding genes harbor multiple

mRNA CS [12], it appears that a “CA” dinucleotide cannot be an absolute requirement for correct cleavage, a situation analogous to the use of both the canonical AAUAAA PAS and its variants by human genes. A functional study supporting this hypothesis demonstrated that the “CA” dinucleotide is preferred, but not required for cleavage machinery recognition, and CS usage at position -1 was found to be in the order of preference A > U > C >> G [5]. This same study and other previous analyses [3, 4] indicated that the CS is located no closer than 10 bases, but no further than 30 bases from the AAUAAA element. Surprisingly, about 11% of CPGs included in our study exhibited a PAS-CS distance greater than 30bp. Thus, we can speculate that certain estimates provided by long-standing polyadenylation studies do not apply to all human transcripts. Nonetheless, the results of the current study must be viewed in the context of two main limitations: a characterization targeting a specific group of genes, and a relatively small number of genes analyzed.

Strikingly, we did not find any reports of a PAS hexamer in about 18% of the CPGs using a reference database (NCBI). For these genes, a computational analysis was developed in this study to identify 3’-most hexamers (putative PAS) in the full corresponding mRNA sequences, because the predominant mRNA sequence is usually the longest one, generated by the 3’-most poly(A) site [12]. Although we did not identify putative PAS for all the referred genes, these novel findings reinforce the relevance of establishing updated methods and/or databases to detect this regulatory element of 3' end processing in human genes. The strategy applied here could be easily employed in similar situations with additional genes outside the CPG context.

In addition, we also explored the frequency of functional APA sites among all CPGs studied. Indeed, APA has emerged as a major player in gene regulation [13] and its pattern in mammals seems to be evolutionarily conserved [36] and regulated in a tissue-specific fashion [37-39]. In this sense, here we reported a strong evidence of APA modulation to the *PTEN* tumor suppressor gene: it contains 61 APA sites differentially used in 22 distinct non-tumoral human tissues obtained from APASdb database. For instance, two of these APA sites are preferentially used in *PTEN* mRNA processing, but their usage quantification was lower in specific tissues, such as kidney and spleen (data not shown). Overall, our analysis using recently released databases (APADB and

APASdb) indicated that approximately 90% of selected CPGs have two or more APA sites. In contrast, a previous analysis estimated that about 54% of human genes are alternatively polyadenylated [12]. Furthermore, in 13 normal human tissues, a polyadenylation sequencing strategy (PA-seq) found that such APA events were present not only in protein-coding genes (38%) but also in noncoding genes (35%) [39], while a genome-wide APA site mapping in some cancer types and tumor cell lines identified around 30% of mRNAs containing APA sites, regardless of the cell type [40]. Taken together, these findings suggest a greater complexity in the regulation of polyadenylation in transcripts specifically derived from CPGs. Importantly, a widespread APA-mediated 3′UTR shortening has been identified across the cancer genome [40-43]. Since shorter 3'UTR isoforms have higher translational efficiency than their longer counterparts due to loss of miRNA regulation, APA events can activate some proto-oncogenes in cancer cells [41, 44]. More recently, Xiang and colleagues (2018) conducted a comprehensive APA characterization in clinical samples comprising 17 tumor types and 739 cancer cell lines, and demonstrated that the complexity of APA profiles might affect clinically actionable genes and drug sensitivity [45].

Lastly, given the scarcity of studies about regulation by miRNAs shared between comprehensive sets of tumor suppressor genes and oncogenes acting on different cell signaling pathways, we investigated whether validated and predicted data sources could identify miRNAs/miRNA families with potential tumor suppressor and oncogenic functions. The mir-192 family was significantly overrepresented as regulator of tumor suppressor genes, having experimentally validated interactions with genes implicated in essential DNA repair pathways, including double-strand break repair by homologous recombination (*BRCA1*, *BRCA2*, and *RAD51*), mismatch repair (*MSH6*), and nucleotide excision repair induced by ultraviolet light (*XPA*). An upregulation of this miRNA family was reported in multiple tumor types including gastric cancer, hepatocellular carcinoma, neuroblastoma, pancreatic ductal adenocarcinoma, and esophageal squamous cell carcinoma [46-50]. Preliminary evidences suggest that this expression pattern could enhance cell proliferation and migration, reduce apoptosis and promote cell cycle progression [46, 47, 50]. Of note, mir-192 was previously identified as an oncomiR in gastric tumorigenesis [51, 52], while its

biological effects in other cancers have been only partially elucidated. These findings support our hypothesis that miRNAs within this family may have a broader oncogenic function. However, mir-192 was found downregulated in colon, colorectal, and lung tumors [53-55]. We also identified some miRNA families that might act as effectors in tumor suppression, including mir-128, mir-1471, mir-483, mir-3170 and mir-218. Recent reports corroborate the assigned role to some of these miRNA families: mir-128 exerts pro-apoptotic effects and it was found as a tumor suppressor in certain tumor types [56-59]; mir-483 suppresses the proliferation of glioma and squamous cell carcinoma cells [60-61]; and emerging tumor-suppressing roles have been described for mir-218 in prostate, breast and lung cancers [62-64]. Up to now, little is known about mir-1471 and mir-3170 roles' in cancer and, if experimentally proven, they might represent novel tumor-suppressive miRNA families. Moreover, mir-34a is a known tumor-suppressive miRNA [28] and it was found regulating several oncogenes included in the present study, validating our *in silico* analysis. Importantly, in a feed-forward loop fashion, mir-34a has a positive effect on p53 transcriptional activity and protein stability (encoded by tumor suppressor gene *TP53*), by targeting multiple p53 inhibitor genes (e.g., the *MDM4* oncogene) [65, 66]. In turn, p53 upregulates this miRNA, contributing to apoptosis and senescence [28, 29]. In fact, *TP53* and *MDM4* genes appeared as direct targets of mir-34a in our analysis. Overall, our results using this computational analysis are in agreement with already available literature data, suggesting that it is a suitable strategy to identify potentially tumor-suppressive and oncogenic miRNAs.

4. Conclusions

In summary, this is the first study to focus on a comprehensive characterization of 3'UTR-related elements (CPE sequences and regulation by miRNAs) in CPGs. This approach provided a landscape of CPE in CPGs which might be useful in the development of molecular analyses covering these frequently neglected 3'UTR regulatory elements. Nevertheless, further studies including additional cancer-related genes should be conducted in order to confirm and/or expand these findings. Additionally, functional validation of

oncogenic and tumor suppressor potential assigned to the miRNA families identified in our *in silico* analysis is warranted.

5. Materials and Methods

5.1. Selection criteria of the cancer predisposition genes (CPGs) included in this study

CPGs included in this study (n=117 genes) were selected from the literature using the following criteria: (1) genes associated with well described genetic syndromes in which the sole or main clinical phenotype is hereditary predisposition to cancer (classic HCS and certain types of familial cancer, n=71); (2) genes associated with syndromes/conditions in which one of the phenotypes is increased tumor predisposition, including some congenital malformation syndromes (n=36); and (3) susceptibility genes to certain sporadic tumor(s), including ones encoding translesion DNA polymerases implicated in cancer either by germline/somatic mutations or by altered expression in tumors (n=10). The minimum requirement needed to link genes and cancer predisposition was a consistent association between sequence variants (or significantly altered expression levels) in such genes and an increased risk for developing certain tumor type(s) through independent case-control studies. Most of the CPGs (n=76) were selected from Garber and Offit review (2005) [67], and others from specific studies detailed in **Table S1**. Based on experimental evidence regarding cancer-associated mutations effects in each gene, CPGs were classified into three categories: oncogenes (n=17), genes whose alterations cause gain-of-function effects to promote carcinogenesis; tumor suppressor genes (n=81), genes in which loss-of-function mutations contribute to the malignant phenotype [68]; and “undetermined function” genes (n=19), which did not fit into the aforementioned categories and/or for which there is insufficient evidence about the mechanism of association with tumor development (main references related to this classification are shown in **Table S1**). In addition, two genes previously reported as putative oncogenes (*SHOC2* e *PALLD*), as well as eight putative tumor suppressor genes (*EPHB2*, *EXT1*,

EXT2, *GPC3*, *ELAC2*, *MSR1*, *POT1* and *PRKAR1A*), were included in the respective groups.

5.2. Polyadenylation database analysis

Briefly, “Nucleotide database” within NCBI (reference source) and a computational method were employed to characterize PAS and CS sequences of the selected CPGs. The recently released APA databases APADB [69] and APASdb [70] were used to assess the occurrence of APA in these genes.

Regarding the NCBI analyses, we searched for the transcript variant with the largest extension in base pairs (bp) derived from each gene, wherein mRNA sequences predicted by computational analysis were not considered. This specific criterion was used to avoid data loss on possible PAS/CS sequences related to a certain gene, which may occur, for example, when the transcript variant has a shorter 3'UTR. After identifying the longest variant transcript, PAS and CS were located in the sequence selecting the features "regulatory" and "polyA_site", respectively. The distance (in bp) between the PAS and CS was calculated according to their positions indicated in the NCBI interface.

Regarding CPGs with no described PAS in the NCBI database, we developed a computational method to identify putative PAS in the complete sequence of corresponding transcripts. First, RefSeq mRNA sequences were obtained from UCSC Genome Browser (human genome version hg19). Screening of the canonical AAUAAA hexamer and its functional variants previously reported in human genome [12] was the second step, restricting the definition of putative PAS to the 3’-most hexamers identified in the transcript sequence. This step was conducted using in-house Perl scripts (further information upon request).

For the APA analyses, APADB data were filtered for CS identified only in human peripheral blood samples and supporting read alignments for each CS were visualized in a genome browser. This database was built by 3’ end sequencing using massive analysis of complementary DNA ends, a high-throughput next-generation sequencing-based technique [69]. Additionally, APASdb allowed the mapped CS counting in all transcript variants for each gene and from 22 normal human tissue samples (searching dataset: hg19

human-all22-tissues). This database, based on the sequencing APA sites (SAPAS) method reported previously [71], provided both the position and usage quantification for a given alternative CS among transcripts derived from the same gene by computing their corresponding normalized-reads [70].

5.3. MicroRNA-mediated regulation analysis

To investigate the miRNAs that may regulate expression of selected tumor suppressor genes (n=81) and oncogenes (n=17), experimentally validated data of miRNA-target gene interactions in humans were collected from the miRTarBase release 6.0 [72], starBase v2.0 [73], TarBase v5.0 [74], and miRecords v4.0 [75] databases. Data derived from miRTarBase was restricted to interactions classified as functional, including those with weak experimental evidence, while starBase data was filtered to keep only interactions predicted by two or more software and supported by at least one experiment ("low stringency" parameter). For other databases, all available interactions were included in the analysis.

Experimental data was complemented by computational target prediction tools, namely using TargetScan v7.2 [76], Diana MicroT-CDS [77], and miRanda-mirSVR (August 2010 Release) [78]. To control false positive rates and restrain the large list of predicted miRNAs, the following filtering criteria were adopted: (a) for TargetScan, only interactions involving 8mer, 7mer and 6mer conserved miRNA sites and with Context++ score lower than -0.2 were considered; (b) for Diana MicroT-CDS, only predictions with a score higher than 0.9 were kept; (c) for miRanda-mirSVR, only interactions with mirSVR score lower than -0.1 (regarded as "good mirSVR score") involving conserved miRNAs were included; (d) the final set of predicted interactions was defined by pairs of miRNA-target genes suggested by at least two computational tools.

Finally, validated and predicted miRNA-target gene interactions collected were filtered in order to identify interactions whose target gene was either a tumor suppressor gene or an oncogene included in this study. Moreover, whenever possible, miRNAs were clustered into families according to the miRBase 21 annotations (some miRNAs do not have a family classification).

These analyses were conducted using in-house R scripts (further information upon request). To better visualize the overlap among validated and predicted data, we created regulatory networks and Cytoscape software version 3.1.0 was used for graphical network visualization [79]. All miRNAs and genes identifiers were mapped to miRBase 21 and HUGO Gene Nomenclature Committee (HGNC) standards, respectively, to maintain uniformity and enable consistent data integration across different sources.

5.4. Statistical analyses

SPSS version 18.0 (IBM) was used for data handling and statistical analyses. Fisher's exact test was employed to compare different PAS hexamer type frequency between oncogenes and tumor suppressor genes. The same test was applied to identify miRNAs and miRNA families overrepresented in the regulation of the groups of oncogenes and tumor suppressors studied, i.e., they have a greater number of target genes among the genes of interest included in this study than would be expected by chance. The complete list of validated and predicted interactions collected from the aforementioned sources, without filtering it for interactions involving all CPGs evaluated in this study, was adopted as the background set in this over-representation analysis. More details about the composition of the background set are detailed in **Table S2**. Statistical tests were two-sided and *P*-values of <0.01 were considered statistically significant.

5.5. Data Availability

The alternative polyadenylation datasets analyzed during the current study are available publicly in the APADB (http://tools.genxpro.net/apadb/) and APASDB (http://genome.bucm.edu.cn/utr/) associated websites. All generated data during this study were included in this published article and their respective Supplementary Information files. The complete raw data from the miRNA-mediated regulation analysis are available separately in Microsoft Excel® spreadsheet file format (.xlsx files).

Competing interests

The authors have no conflicts of interest to declare.

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, grant number 23038.004629/2014-19 to Sandro José de Souza).

Acknowledgments

We are grateful to Clévia Rosset, Isabel Bandeira da Silva, and Bárbara Alemar for their valuable contributions and support. Igor Araujo Vieira and Delva Pereira Leão are postgraduate fellowships recipients from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Vandeclécio Lira da Silva and Marina Roberta Scheid hold postgraduate fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Patricia Ashton-Prolla and Sandro José de Souza are CNPq associated researchers.

References

1. Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev.* **63**, 405-445 (1999).

2. Millevoi, S. & Vagner S. Molecular mechanisms of eukaryotic pre-mRNA 3’ end processing regulation. *Nucleic Acids Res.* **38**, 2757-2774 (2010).

3. Proudfoot, N. J. & Brownlee, G. G. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**, 211-214 (1976).

4. Gil, A. & Proudfoot, N. J. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell* **49**, 399-406 (1987).

5. Chen, F., MacDonald, C. C. & Wilusz, J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.* **23**, 2614-2620 (1995).

6. Legendre, M. & Gautheret, D. Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**, 7 (2003).

7. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.* **9**, 563-576 (2012).

8. Ryan, K., Calvo, O. & Manley, J. L. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**, 565-573 (2004).

9. Keller, W., Bienroth, S., Lang, K. M. & Christofori, G. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. *EMBO J.* **10**, 4241-4249 (1991).

10. MacDonald, C. C., Wilusz, J. & Shenk, T. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol Cell Biol.* **14**, 6647-6654 (1994).

11. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770-1782 (2011).

12. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201-212 (2005).

13. Lutz, C. S. Alternative polyadenylation: a twist on mRNA 3' end formation. *ACS Chem Biol.* **3**, 609-617 (2008).

14. Chen, J. M., Férec, C. & Cooper, D. N. A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. *Hum Genet.* **120**, 1-21 (2006).

15. Michalova, E., Vojtesek, B. & Hrstka, R. Impaired pre-mRNA processing and altered architecture of 3' untranslated regions contribute to the development of human disorders. *Int J Mol Sci.* **14**, 15681-15694 (2013).

16. Hollerer, I., Grund, K., Hentze, M. W. & Kulozik, A. E. mRNA 3'end processing: A tale of the tail reaches the clinic. *EMBO Mol Med.* **6**, 16-26 (2014).

17. Stacey, S. N. *et al*. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet.* **43**, 1098-10103 (2011).

18. Macedo, G. S. *et al*. Rare germline variant (rs78378222) in the TP53 3' UTR: Evidence for a new mechanism of cancer predisposition in Li-Fraumeni syndrome. *Cancer Genet.* **209**, 97-106 (2016).

19. Decorsière, A. *et al*. Decreased efficiency of MSH6 mRNA polyadenylation linked to a 20-base-pair duplication in Lynch syndrome families. *Cell Cycle* **11**, 2578-2580 (2012).

20. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350-355 (2004).

21. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

22. Calin, G. A. *et al*. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* **99**, 15524-15529 (2002).

23. Takamizawa, J. *et al*. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.* **64**, 3753-3756 (2004).

24. Cimmino, A. *et al*. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci USA* **102**, 13944-13949 (2005).

25. Esquela-Kerscher, A. & Slack, F.J. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer* **6**, 259-269 (2006).

26. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001-1010 (2000).

27. Danckwardt, S., Hentze, M. W. & Kulozik, A. E. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.* **27**, 482-498 (2008).

28. Raver-Shapira, N. *et al*. Transcriptional activation of miR-34a contributes to p53-mediated apoptosis. *Mol Cell.* **26**, 731-743 (2007).

29. Navarro, F. & Lieberman, J. miR-34 and p53: New Insights into a Complex Functional Relationship. *PLoS One* **10**, e0132767 (2015).

30. Proudfoot, N. Poly(A) signals. *Cell* **64**, 671-674 (1991).

31. Wickens, M. & Stephenson, P. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3’ end formation. *Science* **226**, 1045-1051 (1984).

32. MacDonald, C. C. & Redondo, J. L. Reexamining the polyadenylation signal: were we wrong about AAUAAA? *Mol Cell Endocrinol.* **190**, 1-8 (2002).

33. Gruber, A. J. *et al*. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**, 1145-1159 (2016).

34. Sheets, M. D., Ogg, S. C. & Wickens, M. P. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* **18**, 5799-5805 (1990).

35. Gehring, N. H. *et al*. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat Genet.* **28**, 389-392 (2001).

36. Ara, T., Lopez, F., Ritchie, W., Benech, P. & Gautheret, D. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* **7**, 189 (2006).

37. Beaudoing, E. & Gautheret, D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res.* **11**, 1520-1526 (2001).

38. Zhang, H., Lee, J. Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol.* **6**, R100 (2005).

39. Ni, T. *et al*. Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* **14**, 615 (2013).

40. Lin, Y. *et al*. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* **40**, 8460-8471 (2012).

41. Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-684 (2009).

42. Lai, D. P. *et al*. Genome-wide profiling of polyadenylation sites reveals a link between selective polyadenylation and cancer metastasis. *Hum Mol Genet.* **24**, 3410-3417 (2015).

43. Erson-Bensan, A. E. & Can, T. Alternative Polyadenylation: Another Foe in Cancer. *Mol Cancer Res.* **14**, 507-517 (2016).

44. An, J., Zhu, X., Wang, H. & Jin, X. A dynamic interplay between alternative polyadenylation and microRNA regulation: implications for cancer (Review). *Int J Oncol.* **43**, 995-1001 (2013).

45. Xiang, Y. *et al*. Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. *J Natl Cancer Inst.* **110**, 379-389 (2018).

46. Feinberg-Gorenshtein, G. *et al*. miR-192 directly binds and regulates Dicer1 expression in neuroblastoma. *PLoS One* **8**, e78713 (2013).

47. Zhao, C. *et al*. Diagnostic and biological significance of microRNA-192 in pancreatic ductal adenocarcinoma. *Oncol Rep.* **30**, 276-284 (2013).

48. Chen, Q. *et al*. Plasma miR-122 and miR-192 as potential novel biomarkers for the early detection of distant metastasis of gastric cancer. *Oncol Rep.* **31**, 1863-1870 (2014).

49. Tan, Y. *et al*. A serum microRNA panel as potential biomarkers for hepatocellular carcinoma related with hepatitis B virus. *PLoS One* **9**, e107986 (2014).

50. Li, S. *et al*. Mir-192 suppresses apoptosis and promotes proliferation in esophageal aquamous cell caicinoma by targeting Bim. *Int J Clin Exp Pathol.* **8**, 8048-8056 (2015).

51. Jin, Z. *et al*. MicroRNA-192 and -215 are upregulated in human gastric cancer in vivo and suppress ALCAM expression in vitro. *Oncogene* **30**, 1577-1585 (2011).

52. Zhang, X. *et al*. Inhibition of the miR-192/215-Rab11-FIP2 axis suppresses human gastric cancer progression. *Cell Death Dis.* **9**, 778 (2018).

53. Feng, S. *et al*. MicroRNA-192 targeting retinoblastoma 1 inhibits cell proliferation and induces cell apoptosis in lung cancer cells. *Nucleic Acids Res.* **39**, 6669-6678 (2011).

54. Karaayvaz, M. *et al*. Prognostic significance of miR-215 in colon cancer. *Clin Colorectal Cancer* **10**, 340-347 (2011).

55. Chiang, Y. *et al*. microRNA-192, -194 and -215 are frequently downregulated in colorectal cancer. *Exp Ther Med.* **3**, 560-566 (2012).

56. Adlakha, Y. K. & Saini, N. miR-128 exerts pro-apoptotic effect in a p53 transcription-dependent and -independent manner via PUMA-Bak axis. *Cell Death Dis.* **4**, e542 (2013).

57. Hauser, B. *et al*. Functions of MiRNA-128 on the regulation of head and neck squamous cell carcinoma growth and apoptosis. *PLoS One* **10**, e0116321 (2015).

58. Wu, L., Shi, B., Huang, K. & Fan, G. MicroRNA-128 suppresses cell growth and metastasis in colorectal carcinoma by targeting IRS1. *Oncol Rep.* **34**, 2797-2805 (2015).

59. Shan, Z. N. *et al*. miR128-1 inhibits the growth of glioblastoma multiforme and glioma stem-like cells via targeting BMI1 and E2F3. *Oncotarget* **7**, 78813-78826 (2016).

60. Wang, L. *et al*. MiR-483-5p suppresses the proliferation of glioma cells via directly targeting ERK1. *FEBS Lett.* **586**, 1312-1317 (2012).

61. Bertero, T. *et al*. Tumor suppressor function of miR-483-3p on squamous cell carcinomas due to its pro-apoptotic properties. *Cell Cycle* **12**, 2183-2193 (2013).

62. Liu, B. *et al*. Tumor-suppressing roles of miR-214 and miR-218 in breast cancer. *Oncol Rep.* **35**, 3178-3184 (2016).

63. Song, L. *et al*. miR-218 suppressed the growth of lung carcinoma by reducing MEF2D expression. *Tumour Biol.* **37**, 2891-2900 (2016).

64. Guan, B. *et al*. Tumor-suppressive microRNA-218 inhibits tumor angiogenesis via targeting the mTOR component RICTOR in prostate cancer. *Oncotarget* **8**, 8162-8172 (2017).

65. Mandke, P. *et al*. MicroRNA-34a modulates MDM4 expression via a target site in the open reading frame. *PLoS One* **7**, e42034 (2012).

66. Okada, N. *et al*. A positive feedback between p53 and miR-34 miRNAs mediates tumor suppression. *Genes Dev.* **28**, 438-450 (2014).

67. Garber, J. E. & Offit, K. Hereditary cancer predisposition syndromes. *J Clin Oncol.* **23**, 276-292 (2005).

68. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

69. Müller, S. *et al*. APADB: a database for alternative polyadenylation and microRNA regulation events. *Database* **pii**; bau076 (2014).

70. You, L. *et al*. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* **43**, D59-67 (2015).

71. Sun, Y., Fu, Y., Li, Y. & Xu, A. Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. *J Mol Cell Biol.* **4**, 352-361 (2012).

72. Chou, C. H. *et al*. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **44**, D239-247 (2016).

73. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, D92-97 (2014).

74. Sethupathy, P., Corda, B. & Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* **12**, 192-197 (2016).

75. Xiao, F. *et al*. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37**, D105-110 (2009).

76. Agarwal, V., Bell, G. W., Nam, J. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).

77. Paraskevopoulou, M. D. *et al*. DIANAmicroT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* **41**, W169-173 (2013).

78. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90 (2010).

79. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.***13**, 2498-2504 (2003).

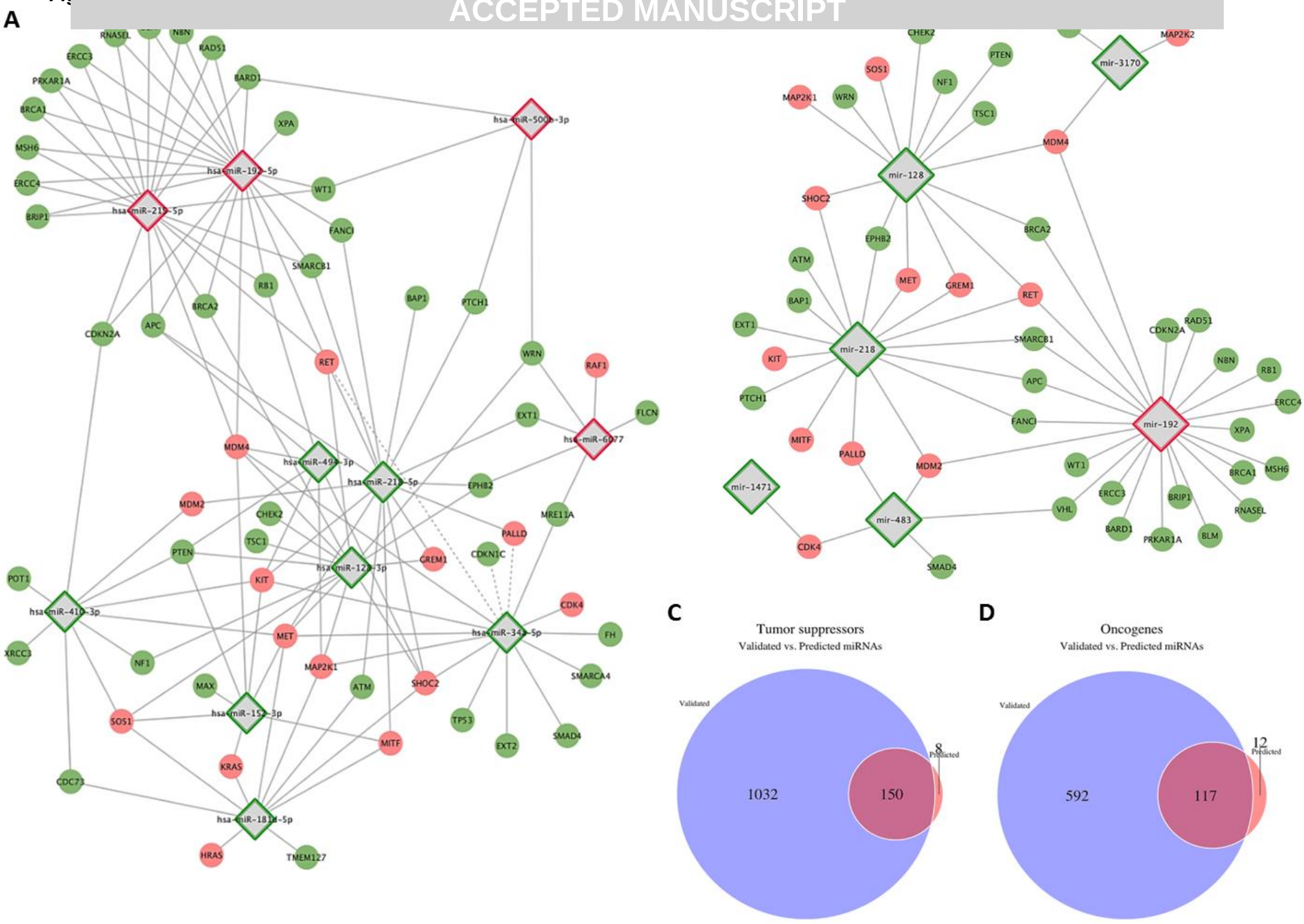
Figures

Fig. 1. APASdb data showing alternative polyadenylation/cleavage sites (CS) and polyadenylation signals (PAS) mapped to the transcript derived from *PTEN* gene, containing the 5' and 3' flanking region of 1kb. This database identified 61 differentially used alternative CS in the *PTEN* mRNA processing among 22 normal human tissues (Mean reads: 75, searching dataset: hg19 human-all22-tissues). Position at chromosome 10 (mentioned after “pA” code), PAS sequences and usage quantification (%) are indicated for each CS. As shown in the original APASdb output, the first eleven *PTEN* polyadenylation sites have their usage quantification depicted by colorful bar charts, while the other sites appear in gray. Of note, two CS located in coding regions (chr10:89726131 and chr10:89726870) had higher usage quantification and were highlighted by rectangles with dashed lines.

Fig. 2. Regulatory networks and Venn diagrams obtained by combining validated and predicted interactions between miRNAs and tumor suppressor genes/oncogenes. miRNAs (A) and miRNA families (B) significantly overrepresented ($P<0.01$) as regulators (diamond nodes) of expression between the groups of tumor suppressor genes (green circle nodes) and oncogenes studied (red circle nodes) are shown in these interactomes. Red-outlined diamonds indicate miRNAs with potential tumor suppressor function, whereas the green-outlined ones represent potentially oncogenic miRNAs. Solid lines denote experimentally validated interactions and dashed lines are high confidence computationally predicted interactions. Interactome (A) was built considering only miRNAs with at least three target genes among tumor suppressor genes and oncogenes. Intersections among the sets of validated and predicted miRNAs found to modulate expression of tumor suppressor genes (C) and oncogenes (D) are indicated in Venn diagrams.

Fig. 1.

Fig



Highlights

- There are cancer predisposition genes (CPGs) without defined poly(A) signals;
- Optimal/preferential cleavage site sequences usually are not used by most CPGs;
- The majority of CPGs have alternative polyadenylation sites;
- *In silico* analysis suggests a broader oncogenic function for the mir-192 family.

Abbreviations

CPE: core polyadenylation elements; mRNAs: messenger RNAs; PAS: polyadenylation signal; CS: cleavage site; CPGs: cancer predisposition genes; miRNAs: microRNAs; NCBI: National Center for Biotechnology Information; nt: nucleotides; CPSF: polyadenylation specificity factor; CstF: cleavage stimulatory factor; CFI/CFII: cleavage factors I and II; PAP: poly(A) polymerase; APA: alternative polyadenylation; 3'UTR: 3' untranslated region; HCS: hereditary cancer syndromes; ONC: oncogene; TSG: tumor suppressor gene; bp: base pairs

Author contributions statement

Conceptualization, Igor Araujo Vieira, Mariana Recamonde-Mendoza, Delva Pereira Leão, Sandro José de Souza and Patricia Ashton-Prolla; Funding acquisition, Sandro José de Souza and Patricia Ashton-Prolla; Methodology, Igor Araujo Vieira, Vandeclecio Lira da Silva, Delva Pereira Leão and Marina Roberta Scheid; Project administration, Patricia Ashton-Prolla; Supervision, Sandro José de Souza and Patricia Ashton-Prolla; Writing – original draft, Igor Araujo Vieira, Mariana Recamonde-Mendoza, Vandeclecio Lira da Silva and Patricia Ashton-Prolla; Writing – review & editing, Igor Araujo Vieira, Mariana Recamonde-Mendoza, Vandeclecio Lira da Silva, Sandro José de Souza and Patricia Ashton-Prolla.