Universidade de Aveiro Departamento de Química
Ano 2013/14

Artur Filipe Cardoso Duarte Rodrigues

Implications of A to I RNA editing in circular RNA biogenesis

Universidade de Aveiro Departamento de Química
Ano 2013/14

Artur Filipe Cardoso Duarte Rodrigues

Implicações da edição de ARN do tipo A para I em ARN circular

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biotecnologia, ramo de Biotecnologia Molecular, realizada sob a orientação científica do Doutor Alekos Athanasiadis, Investigador Principal do grupo "Interações Proteínas – Ácidos Nucleicos", no Instituto Gulbenkian de Ciência e do Professor Doutor Manuel Santos, Professor Associado do Departamento de Biologia da Universidade de Aveiro.

Ao meu avô.

o júri

presidente

Prof. Doutora Etelvina Maria de Almeida Paula Figueira
Professora Auxiliar do Departamento de Biologia da Universidade de Aveiro

Doutor José Bártholo Pereira Leal
Investigador Principal do grupo “Genómica Computacional” no Instituto Gulbenkian de Ciência

Doutor Alekos Athanasiadis
Investigador Principal do grupo “Interações Proteínas – Ácidos Nucleicos” no Instituto Gulbenkian de Ciência

agradecimentos

First of all, I would like to thank my supervisors, Doctor Alekos Athanasiadis and Professor Manuel Santos, for their guidance and the opportunity to work at Instituto Gulbenkian de Ciência. Working at an institute as the IGC, where I had the opportunity to meet remarkable scientists from all over the world and discuss science. This was really important for me to understand science, and its international, collaborative environment fascinated me every single day.

When I arrived at the IGC, I also found an amazing group that made me feel very welcomed. I want to thank everyone who joined me over the last year – Diogo, Krzysztof, Luísa, Matteo, Klito, Nuno and Daniela. I will definitely miss some of the good times we spent together! I want to thank especially Diogo and Krzysztof for their patience to help me surpassing my inexperience in programming and for their critical reviews which were really helpful to improve the quality and effectiveness of my work.

Como esta dissertação marca o fim de um ciclo de estudos importante, queria agradecer a todos os que me acompanharam no meu caminho:

- Aos meus professores de Biologia na Escola Secundária Homem Cristo, Armindo Jorge Fernandes e Rosa Amado, pela forma como me cativaram para esta área de estudo, o que me levou a ingressar no curso de Biotecnologia na Universidade de Aveiro;

- Aos meus colegas do Departamento de Química e da Tuna Universitária de Aveiro, pela amizade e experiências vividas nestes anos académicos, onde sem dúvida pude crescer imenso;

- A todos os meus amigos de longa data, em particular ao Pedro e ao João, por todos os anos de amizade, conselhos e boas vivências;

- Ao Professor Armando Silvestre, à Doutora Carmen Freire e ao Nuno Silva, por me terem ajudado a iniciar o meu percurso científico e me ensinarem a ser um cientista;

- À Bárbara, pelo teu amor, amizade e carinho que tanto me apoiaram e fizeram cada vez acreditar mais em mim e em nós, e em como podemos lutar para atingir os nossos sonhos, lado a lado;

- À minha família. Pela educação e transmissão dos valores que me moldaram no homem que sou agora, com momentos bons e menos bons, que nos fazem apreciar realmente o que é preciso valorizar na vida.

"Para ser grande, sê inteiro: nada
Teu exagera ou exclui.
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, porque alta vive."
- Ricardo Reis

palavras-chave

ARN circular, edição de ARN, elementos Alu, splicing alternativo, bioinformática.

resumo

Os ARN circulares (circRNAs) foram identificados como novos padrões de *splicing* alternativo que emergiram recentemente como uma configuração naturalmente abundante, conservada em *Eukarya*, *Bacteria* and *Archaea*. Foi demonstrado que os circRNAs são enriquecidos em elementos Alu nas suas regiões flanqueadoras, que podem formar pares com outros elementos em orientação inversa nos flancos opostos. Assim, postulou-se que esse emparelhamento poderia promover a circularização do ARN ao aproximar ambos os *splice sites*.

Elementos Alu são retrotransposões específicos nos genomas dos primatas e de natureza repetitiva, que pertencem à família dos SINEs e constituem cerca de 10% do genoma humano. A abundância de possíveis emparelhamentos entre elementos Alu origina substratos estáveis que podem ser alvo de edição de ARN do tipo A para I. Este fenómeno consiste numa modificação pós-transcricional, em que os nucleótidos adenosina (A) são convertidos em inosina (I), que são interpretados como guanosinas pela maquinaria celular, com implicações no splicing alternativo.

O objetivo desta tese consiste em entender a influência da edição do ARN do tipo A para I nos elementos Alu invertidos que flanqueiam os circRNAs, através de análise computacional de dados relativos a circRNAs publicamente disponibilizados. Confirmámos a nossa hipótese de que a edição do ARN é reduzida nestes elementos Alu, confirmando a sua importância na biogénese dos circRNAs.

keywords

Circular RNA, RNA editing, Alu elements, alternative splicing, bioinformatics.

abstract

Circular RNAs (circRNAs) have been stated as new splicing patterns which have emerged recently as a naturally abundant configuration, conserved in *Eukarya*, *Bacteria* and *Archaea*. CircRNAs were shown to be enriched in Alu elements in their flanking regions, which may form pairs with other repeats in inverted orientation in the opposite flank. Therefore, it has been postulated that pairing between inverted Alu elements may promote RNA circularisation by bringing closer both splice sites.

Alu elements are repetitive, primate-specific retrotransposons from the SINE family, which comprise about 10% of the human genome. Abundance of inverted Alu pairs creates stable substrates for A to I RNA editing. A to I RNA editing is a post-transcriptional modification, where adenosines (A) are converted into inosines (I), which are interpreted as guanosines by the cellular machinery, with implications on alternative splicing.

In this thesis, we aimed to understand the influence of A to I RNA editing in inverted Alu elements flanking circRNAs through computational analysis of publicly available circRNA datasets. We hypothesised and confirmed that A to I RNA editing is reduced in these Alu elements, confirming their importance in circRNA biogenesis.

ABBREVIATIONS

5-HT$_{2C}$	5-hydroxytryptamine (serotonin) receptor 2C
A	adenosine
ADAR	adenosine deaminase that acts on ribonucleic acids
ADAT	adenosine deaminase that acts on transfer ribonucleic acids
APOBEC	apolipoprotein B mRNA editing enzymes, catalytic polypeptide-like
bp	base pair
C	cytidine
cDNA	complementary deoxyribonucleic acid
circRNA	circular ribonucleic acid
ciRNA	circular intronic long noncoding ribonucleic acid
CLIP	crosslinking followed by immunoprecipitation
DAVID	Database for Annotation, Visualisation and Integrated Discovery
DNA	deoxyribonucleic acid
dsRBD	double-stranded ribonucleic acid binding domain
dsRNA	double-stranded ribonucleic acid
EMBOSS	European Molecular Biology Open Software Suite
EST	expressed sequence tag
G	guanosine
GABA$_A$	γ-aminobutyric acid receptor A
GluR-B	glutamate receptor B
GUI	graphical user interface
hnRNP	heterogeneous nuclear ribonucleoprotein
I	inosine
IRES	internal ribosome entry site
KEGG	Kyoto Encyclopedia of Genes and Genomes
kb	kilobase (1000 bp)
kDa	kilodalton
mRNA	messenger ribonucleic acid
miRNA	micro ribonucleic acid
MS	mass spectrometry
ORF	open reading frame

pre-mRNA	precursor messenger ribonucleic acid
RBP	ribonucleic acid binding protein
RISC	ribonucleic acid-induced silencing complex
RNA	ribonucleic acid
RNAi	ribonucleic acid interference
RNP	ribonucleoprotein
rRNA	ribosomal ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
SINE	short interspersed element
siRNA	small interfering ribonucleic acid
SNP	single nucleotide polymorphism
SR	serine/arginine-rich protein (splicing)
SS	splice site
ssRNA	single-stranded ribonucleic acid
tRNA	transfer ribonucleic acid
U	uridine
UTR	untranslated region
UV	ultraviolet radiation

INDEX

1. INTRODUCTION

Since the publication of the first draft sequence of the human genome, substantial progress has been made in human genetics and genomics research [1]. It was discovered a surprisingly small number of about 26000 protein-coding genes for the predicted plethora of proteins which confer the molecular complexity of our species [2, 3], even though recent studies have suggested that protein diversity may be much lower than expected [4]. According to the central dogma of molecular biology, genetic information flows from deoxyribonucleic acid (DNA) to a complementary copy of ribonucleic acid (RNA) through transcription, ending with its translation to a protein [5]. The fact that RNA may be subject to several modifications on its sequence suggests its central role on proteome diversity.

Phenotypic complexity has been associated to multiple protein variants, derived from mechanisms such as alternative splicing which occur on precursor messenger RNA (pre-mRNA) [6, 7]. Moreover, most RNAs are noncoding and possess different regulatory functions, not only to control gene expression but also to be targeted by proteins [1].

Some enzymes can alter the coded information by catalysing nucleic acid sequence rearrangements or altering single nucleotides. Sequence is not fate.

1.1. RNA EDITING

Traditionally, messenger RNA (mRNA) undergoes various post-transcriptional modifications, which include 5'-capping, polyadenylation at the 3'-end, and alternative splicing. These modifications convert RNA precursors to mature RNA before translation [8].

Unlike other post-transcriptional modifications, RNA editing involves mainly the conversion of individual bases and the insertion or deletion of homonucleotide runs. RNA editing was first described in trypanosome mitochondria, where polyuridine sequences were inserted into RNA [9]. The most common types of substitutional RNA editing consist in the deamination of adenosine (A) bases to inosine (I) (**Figure 1**), by adenosine deaminases acting on RNA (ADARs), and the conversion of cytidine (C) bases to uridine (U) (**Figure 2**), by apolipoprotein B mRNA editing enzymes, catalytic polypeptide-like (APOBECs) and other cytidine deaminases [10, 11].

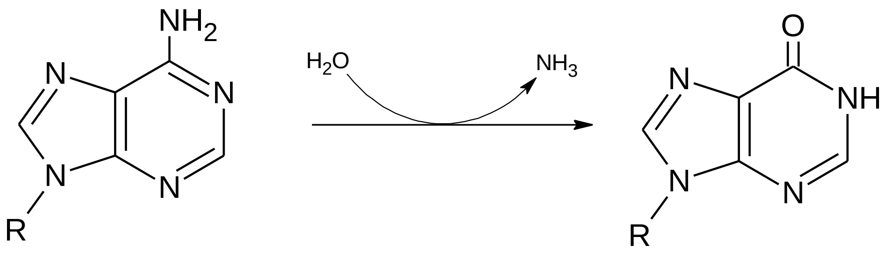


Figure 1. Hydrolytic deamination of adenosine (A; left) to inosine (I; right). R stands for a ribose bound to a 5'- and a 3'-phosphate, in the RNA backbone.

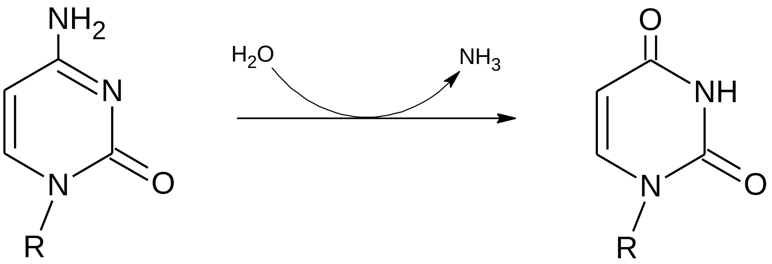
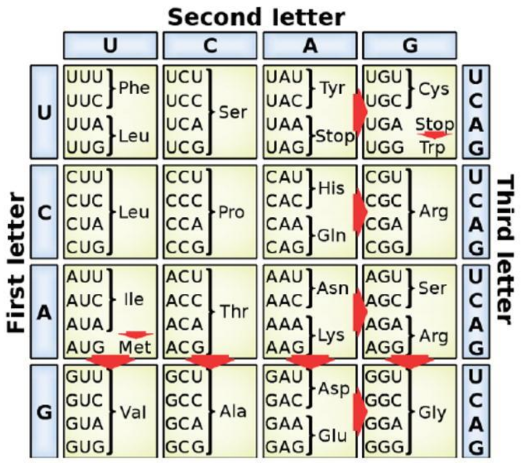


Figure 2. Hydrolytic deamination of cytosine (C; left) to uridine (U; right). R stands for a ribose bound to a 5'- and a 3'-phosphate, in the RNA backbone.

Both of these mechanisms may result in changes of coding sequences in mRNA and therefore they may alter the corresponding amino acid sequence and protein function, relatively to the originally designated by DNA [12]. Editing events consisting of insertions or deletions may cause frameshift mutations and create new open reading frames (ORFs).

A to I RNA editing can generate new start codons or destroy existing ones, and prevent nonsense mutations by removing stop codons and converting them to tryptophan (**Figure 3**).



Second letter

First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 3. Amino acid conversions promoted by A to I editing [12].

Moreover, A to I RNA editing may modify splice sites (SS) and branch points, creating new splice patterns (alternative splicing) and providing different combinations of processed mRNAs [12]. For example, AU and AA dinucleotides can be converted to IU and AI, which are recognised by the spliceosome as the canonical GU and AG typically found at 5'-SS and 3'-SS, respectively (**Figure 4**). In addition, the branch point may be modified, preventing its recognition by the spliceosome.

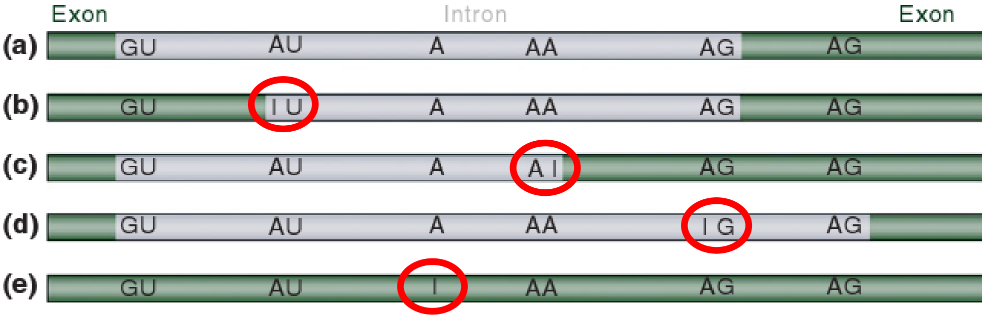


Figure 4. Influence of A to I editing on alternative splicing. The original splice pattern (a) may be changed through the creation of new splice sites (b, c), or disruption of existing ones (d) or branch points (e). Adapted from [12].

RNA editing at the 5' and 3' untranslated regions (UTRs) is more common and may affect mRNA stability and processing, which may influence translation [8, 10]. However, translational efficiency is affected more preponderantly by the nature of double-stranded UTRs than editing [13].

Although mRNAs have been more described as associated with editing processes, other types of noncoding RNA such as transfer RNA (tRNA), ribosomal RNA (rRNA) and 7SL RNA may also be edited.

Editing of tRNA is mediated by adenosine deaminases that act on tRNA (ADATs), which are present in all known living organisms and have been hypothesised as the evolutionary ancestors of ADARs. Changes in tRNA primary sequence may develop new mismatches and correct previous ones, regulating tRNA stability. Modifications at the D and T loops influence tRNA tertiary structure and their affinity to aminoacyl-tRNA synthetases and ribosomes. These enzymes proofread and catalyse aminoacylation of the cognate tRNA at the 3'-CCA end of its acceptor stem, after which they are delivered to the A (aminoacyl) site of the ribosome for translation [14].

In addition, RNA editing at the anticodon loop may alter recognition of mRNA codons, as a result of typical deamination reactions catalysed by ADAT in a variety of eukaryotic species [8, 15]. These modifications stabilise and restrict the conformation of the anticodon domain, preserving the open loop structure during the acquisition of the tertiary structure which is required for codon binding (**Figure 5**) [16, 17].

Modifications of rRNA are also functionally important, because they are more frequent at conserved regions which are responsible for tRNA selection and proofreading, influencing translational efficiency [8].

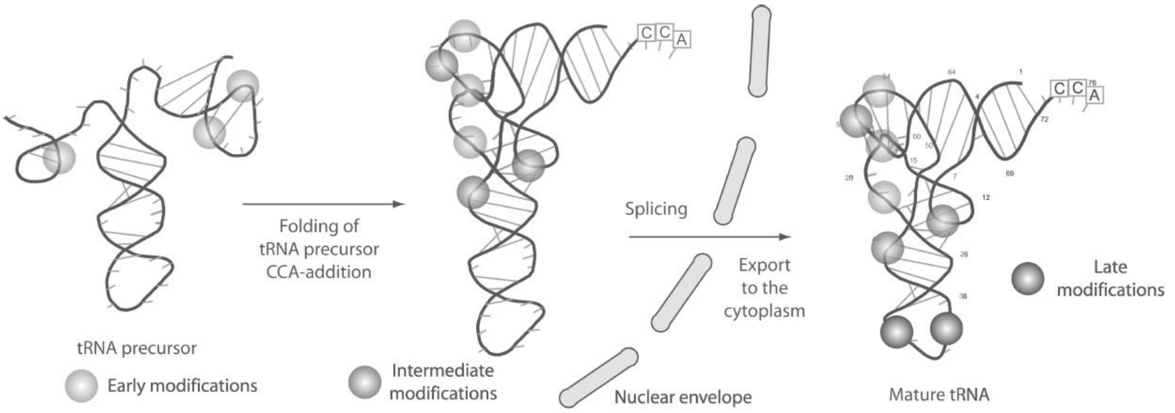


Figure 5. tRNA processing through several post-transcriptional modifications [17].

RNA editing appears as a controlled alternative to gene mutations, providing different combinations of transcripts with different extents of editing [18].

1.1.1. Adenosine Deaminases that Act on RNA (ADARs)

In the most common type of RNA editing in animals, adenosine is converted into inosine which is read as guanosine (G) by the ribosomes during translation. Reverse transcription assays proved that inosine leads to the integration of cytosine in the complementary DNA (cDNA) strand (**Figure 6**) [8, 10].

ADARs have been cloned and characterised in several animal species [8, 10]. Three main types of ADARs have been described and conserved in vertebrates (ADAR1, ADAR2 and ADAR3), with the same substrate specificities and similar activity to ADATs [10, 19].

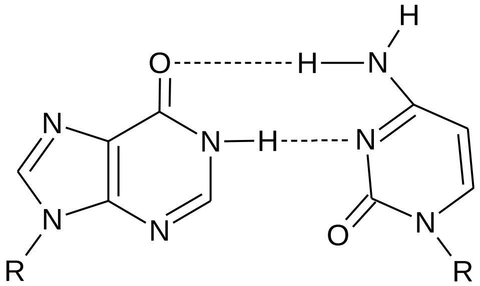


Figure 6. Wobble base pairing between I (left) and C (right).

ADAR1 and ADAR2 are expressed in most human tissues and form homodimers [20, 21], whereas ADAR3 only exists as a monomer in post-mitotic cells, located in certain parts of the central nervous system such as thalamus and amygdala [22].

Dimerisation appears to be necessary for the catalytic deaminase activity, and may explain why ADAR3 is not active [20]. Ensterö *et al.* suggested that one monomer is required to perform the catalytic activity on double-stranded RNA (dsRNA), while the other stabilises the enzyme during the reaction [23].

As a monomer, ADAR3 appears to decrease the efficiency of the other two isoenzymes through binding to potential substrates without editing them, or possibly through dimerisation with ADAR1 or ADAR2 monomers [20, 22].

1.1.1.1. ADAR STRUCTURE AND CATALYTIC ACTIVITY

All ADARs contain a highly conserved catalytic deaminase domain in their C-terminal region and a variable number of double-stranded RNA-binding domains (dsRBDs) in their N-terminal region (**Figure 7**).

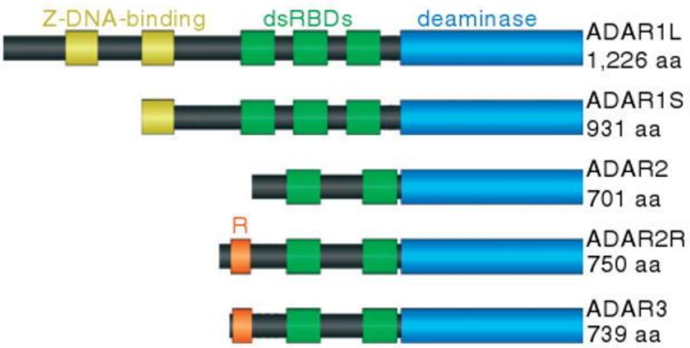


Figure 7. Domains described in all human ADAR isoforms [12].

The catalytic domain resembles the same one existing in cytidine deaminases such as APOBEC. Its center consists of a zinc atom that is coordinated by multiple residues, which promotes the nucleophilic attack to the C6 carbon atom from adenosine, during the hydrolytic deamination reaction [8]. The zinc center is located in a deep pocket in the enzyme surface, which is surrounded by electrostatic potential that promotes dsRNA binding. Macbeth *et al.* demonstrated that inositol hexakiphosphate acts as a cofactor that is required to stabilise the catalytic center folding, and consequently modulates editing activity [24].

ADAR dsRBDs present α-β-β-β-α topology, in which the two α-helices are packed against three anti-parallel β-sheets [12, 19]. The N-terminal α1-helix and the loop between β1 and β2 interact with the RNA minor groove; whereas the short loop between β3 and the C-terminal α2-helix binds to the dsRNA backbone, across the major groove (**Figure 8**) [25]. ADAR isoforms contrast essentially in the number and spacing of their dsRBDs, which are expected to increase the affinity for dsRNA [12]. Nevertheless, ADAR1 dsRBD2 appears to be dispensable for dsRNA-binding and deaminase activities in A to I editing. ADAR1 dsRBD1 seems to bind and direct the RNA substrate to the catalytic center, while dsRBD3 is necessary for deaminase activity [26].

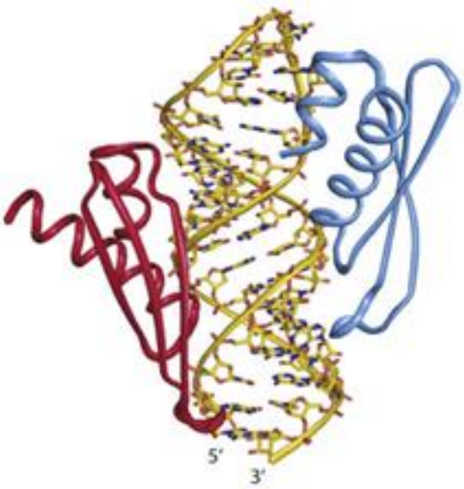


Figure 8. Binding of two dsRBDs to a stem-loop RNA [27].

ADAR1 contains three dsRBDs and exists in two isoforms which result from alternative splicing of an upstream exon, which is skipped to a downstream methionine [19]. Constitutive ADAR1 (ADAR1S; **Figure 7**) is a 110-kilodalton (kDa) isoform, which contains a single Z-DNA binding domain (Zβ). The presence of interferon induces synthesis of a longer 150-kDa isoform (ADAR1L; **Figure 7**), which includes an additional domain (Zα) with a nuclear export signal, which allows ADAR1L to translocate from the nucleus to the cytoplasm [10]. Although ADAR1S is predominantly located in the nucleus and lacks a nuclear export signal, it may also shuttle between the nucleus and the cytoplasm due to the interaction of a nuclear localisation signal around both N- and C- terminal regions to the third dsRBD with import factor transportin-1 and export factor exportin-5 [28, 29]. This interaction appears to depend on dsRNA binding. In the absence of dsRNA, this dsRBD acts as a scaffold that modulates binding to transportin-1 and, hence, the translocation of ADAR1 to the nucleus. On the other hand, the third dsRBD does not bind to transportin-1 in the presence of dsRNA, preventing ADAR1 from carrying dsRNA back to the nucleus by maintaining ADAR1 in the cytoplasm [29].

Both Zα and Zβ binding domains present similar helix-turn-helix β-sheet folding between the first three α-helices and β-sheets, constituting α-β-α-α-β-β topology. However, Zβ presents an extra α-helix packed against the core folding at the C-terminal, establishing α-β-α-α-β-β-α topology (**Figure 9a**) [30]. Despite this addition, allied to the fact that Zβ lacks several residues which are essential to bind Z-DNA, this domain is highly conserved in ADAR1 among several species, which suggest that Zα and Zβ probably execute different functions [19].

Zα domain is the main responsible for binding to both left-handed double helical nucleic acids (Z-DNA and Z-RNA) [30], suggesting the importance of conformation and shape of the zigzag backbone (**Figure 9b**), rather than sequence [19].

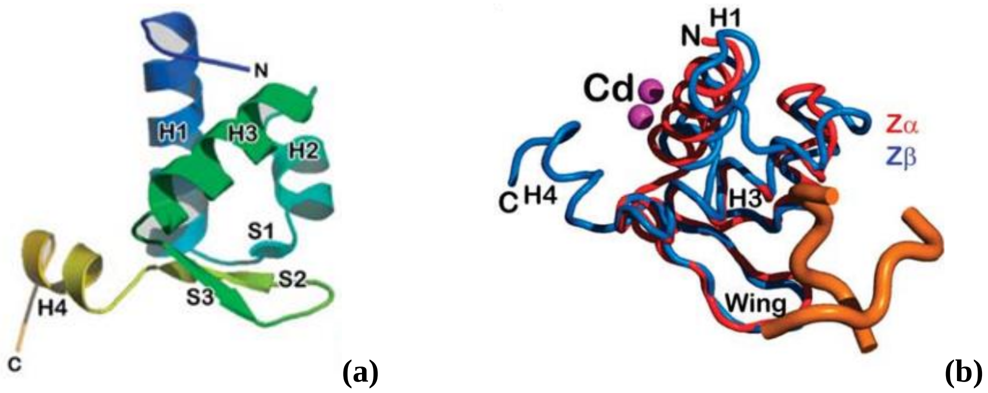


Figure 9. Topology of ADAR1 Zβ domain, where α-helices are presented as H and β-sheets are showed as S **(a)**, and comparison with Zα structure relatively to the backbone of Z-DNA, displayed in orange **(b)** [30].

ADAR2 is the most studied isoenzyme and it is conserved in a variety of eukaryotic organisms (**Figure 13**). Some species only express ADAR2, which is localised mostly in the nucleolus [31]. Despite it has only two dsRBDs, ADAR2 is essential to promote an efficient A to I RNA editing in the nucleus [32]. While the first domain seems to bind specifically to dsRNA, the second domain appears to move the deaminase domain towards the editing site, without blocking the access of adenosine to the catalytic domain [33]. Considering their functional activity, the first dsRBD of ADAR2 corresponds to dsRBD1 of ADAR1, while the second dsRBD of ADAR2 resembles dsRBD3 of ADAR1.

A minor splicing variant of ADAR2 presents an additional upstream exon that extends the coded protein to a longer isoform (ADAR2R). This additional exon encodes a domain very similar to the arginine-rich R-domain in ADAR3, which allows it to bind to single-stranded RNA (ssRNA), cooperating with dsRBDs that bind to dsRNA [10, 19, 22]. Therefore, binding to ssRNA may avoid formation of secondary structures and reduce the availability of dsRNA, which is the main substrate of ADAR1 and ADAR2 [12].

1.1.1.2. ADAR SPECIFICITY AND AFFINITY FOR DSRNA SUBSTRATES

The most representative substrates for ADAR consist mainly of long, unbranched double helices, rather than short double helices branching off from non-helical sequences, which are typical in tRNA and rRNA [10]. ADAR dsRBDs recognise dsRNA structure, preferring A-form helices and stem-loop structures [33]. Most ADAR substrates are intramolecular hairpins resulting from natural backfolding events in a single RNA molecule. Nevertheless, some substrates may be formed through intermolecular interactions between an mRNA molecule with its naturally occurring antisense chain during transcription [10].

The binding affinity of dsRBDs to the substrate dsRNA varies substantially with the bases' stereochemical characteristics [19]. ADAR2 is associated with site-selective editing, whereas ADAR1 is more prone to promiscuous hyper-editing, possibly due to the additional dsRBD. After mismatches and bulges are introduced in long duplexes by ADAR1, dsRNA is divided into smaller regions that may be edited by ADAR2, which is fixed in a specific position so that the deaminase domain approach the targeted adenosine [34].

Stability of dsRNA is essential for a higher editing efficiency. Long, perfect dsRNA substrates with over 50 base pairs (bp) may suffer promiscuous editing of up to 50% of all adenosines, creating I-U mismatches [10]. These mismatches result in internal loops, which multiply during the deamination reaction and gradually decrease stability, due to electrostatic repulsion between adjacent phosphate groups, without the base-base stacking interactions. A sufficiently high number of mismatches results in unwinding dsRNA to ssRNA. On the other hand, editing in A-C mismatches may correct them to stable I-C pairs, which can help increase RNA stability.

ADAR selectivity for editing sites increases with the number of loops in dsRNA, as ADARs tend to detach from the substrate and stops the reaction after fewer deaminations. ADAR1 can target loops with more than six nucleotides, which have the necessary length to uncouple the helix from adjacent double-stranded regions [35]. Tian *et al.* showed that A to I editing in animals can occur in dsRNAs as short as 10 bp, which reflects the possibility of ADARs to interfere with the production of small interfering nuclear RNA (siRNA) [36].

Therefore, RNA editing is more selective in short dsRNAs with imperfect base pairing, bulges and loops, as those structures are less stable and tend to lose the double-stranded character more easily [10, 12] (**Figure 10**).

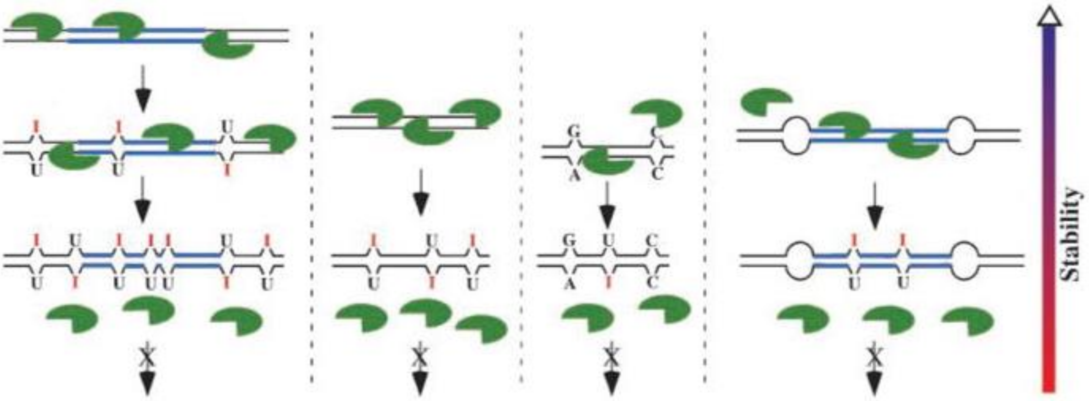


Figure 10. ADARs (green) editing dsRNAs of differing stabilities [10]. The sequence is modified more selectively when placed between internal loops, which have lower stability.

Tertiary structure is decisive on editing selectivity and influences conservation of specific sites where the same adenosines are edited with higher efficiency, especially in repetitive elements [23, 34]. Rieder *et al.* demonstrated the influence of accessory RNA duplexes in editing a target duplex and the importance of conserved tertiary structures to stabilise and direct ADARs to selectively deaminate adenosines [37]. Therefore, structure is essential for substrate recognition and editing.

Moreover, location of adenosines influences its probability of editing, since terminal regions are less prone to be edited [38, 39], possibly due to the length and the relative position of α1 helix to the dsRBD fold [33]. ADAR1 dsRBDs present longer α1 helices and lack the ADAR2 region involved in sequence-specific interactions, which may explain different substrate specificities [19]. In addition, after editing of the first nucleotide, other editing sites mostly occur separated by a minimum distance of 10-12 nucleotides from each other and from the initiation site [34].

However, ADAR dsRBDs do not only recognise secondary structures. Whereas the positively charged N-terminal of α2 helix interacts with the non-bridging oxygen of the phosphodiester bond in the major groove of the substrate's backbone, the α1 helix and the β1-β2 loop promote sequence-specific contacts with the 2'-OH groups of the ribose sugar rings in the minor groove [19].

Kuttan *et al.* suggest a neighboring preference for target adenosines, which influences its ability to flip and expose the base to the catalytic center for deamination [40]. Stereochemical limitations between protein side chains and nucleoside bases in the minor groove help discriminate specific sequences, through steric clashes which may arise after nucleotide-flipping of the substrate (**Figure 11**) [19].

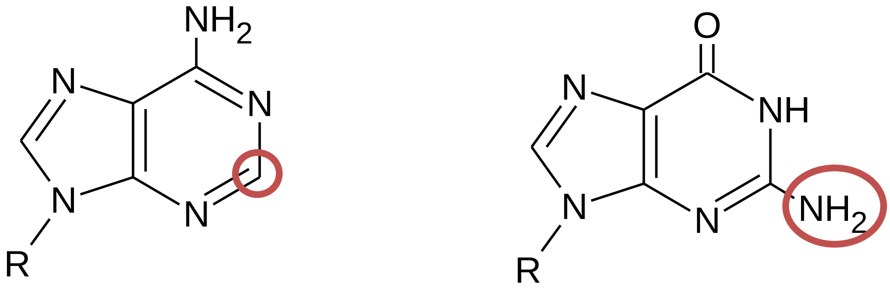


Figure 11. Difference at the position 2 of the purine rings between A and G, marked with red circles. The 2'-H proton of A is non-polar and small, allowing it to accommodate hydrophobic side chains in its close vicinity, while the 2'-NH_2 group of G is a polar hydrogen bond donor, which interacts preferentially with hydrophilic side chains.

Furthermore, the close neighborhood may affect adenosine flipping and exposure to the catalytic domain of ADARs. Although ADARs prefer stable secondary structures, these should not be too stable in order to allow nucleotide flipping. Therefore, ADARs may have sequence preferences relatively to this neighborhood [41].

Human ADAR1 and ADAR2 do not have significant differences regarding which adenosines they target within dsRNA. Even though they share the same preference at the 5'-nearest neighbor (U > A > C > G), ADAR1 and ADAR2 appear to demonstrate slightly different preferences in the 3'-neighborhood (G > C ≈ A > U for ADAR1; G > C > U ≈ A for ADAR2) [42]. Therefore, there are some preferred triplets for ADARs (**Figure 12**) such as UAG, UAC, AAG and AAC, which may contribute to a higher affinity of ADAR domains and confer a higher efficiency, representing a 30-50% conversion from A to I, depending on the mismatch stability [10].

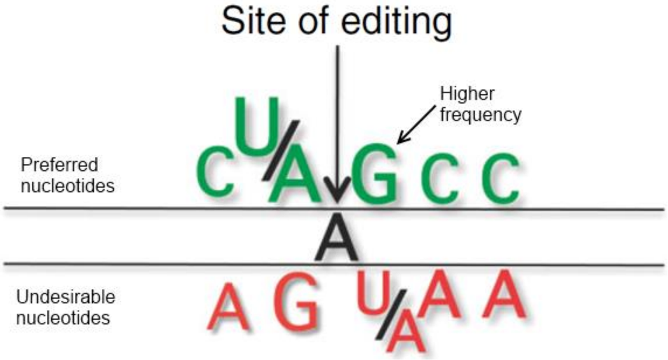


Figure 12. Sequence preference for A-to-I editing. Adapted from [34].

The need of sufficiently stable secondary structures reflect the preference for G and C in the closest neighborhood. However, the preferred neighbor immediately upstream of the editing site is U or A, which may facilitate adenosine flipping and consequent editing [43]. Moreover, edited adenosines tend to be paired with cytosines, forming less stable bonds which favor flipping [41].

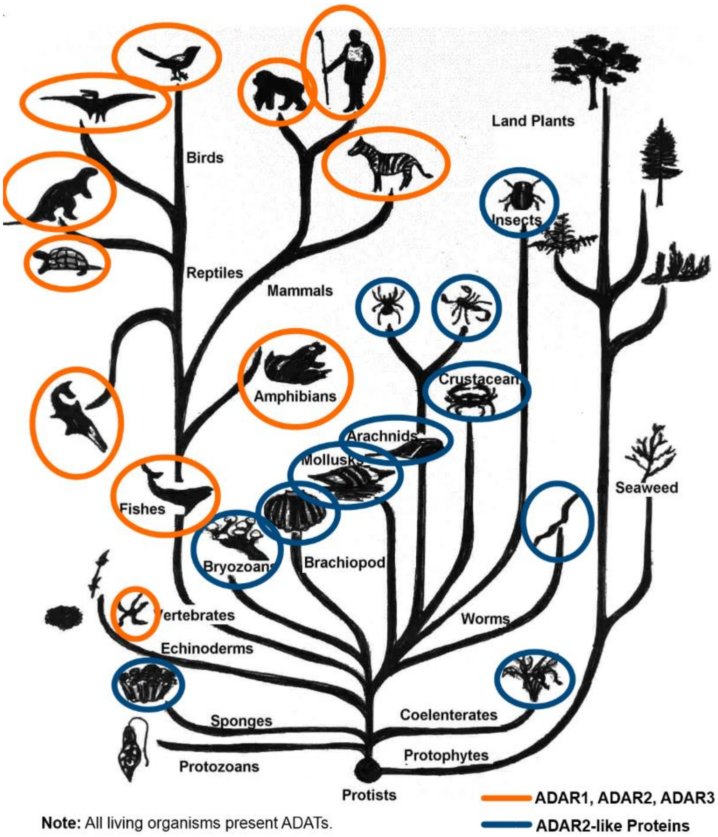
Increasing knowledge of ADAR specificity allowed the detection of A to I editing sites. Due to technological limitations, the first attempts to detect editing sites were based on the comparison of clusters of A to G mismatches, found in cDNA and expressed sequence tags (ESTs), with genomic RefSeq sequences [44, 45]. The great majority of editing sites was found in Alu elements [44].

Recent advances in RNA sequencing technologies allowed the identification of a greater number of editing sites. However, high throughput sequencing still has some limitations on the library construction, mostly due to artifacts derived from incorrect mapping of sequencing reads which may be too short to align and provide the correct location in the genome, and may also contain single nucleotide polymorphisms (SNPs) that can be mistakenly interpreted as editing sites and result in false positives [46, 47].

More recent methodologies try to address these limitations through statistical analysis to filter out SNPs in the human genome [47, 48] and redundant reads corresponding to the same positions [49]. With the increasing information on editing sites, it will be possible to predict and comprehend ADAR activity more thoroughly [50].

1.1.2. Biological relevance and evolution of ADARs and RNA editing

RNA editing is widespread in several taxa, with essential functions. A to I RNA editing occurs in Metazoa (**Figure 13**), with a high level of conservation of ADAR [32]. However, only a small fraction of editing sites is conserved across mammals [51]. ADARs are especially important in the central nervous system of *Drosophila melanogaster* [32, 52–55] and *Caenorhabditis elegans* [56].



Note: All living organisms present ADATs.

Figure 13. Presence of ADAR in *Eukarya*. Adapted from [57].

Editing is essential in the central nervous system of mammals, influencing the expression of several neurotransmitter receptors, such as glutamate receptor B (GluR-B) [58, 59], γ-aminobutyric acid receptor ($GABA_A$) [60] and serotonin 5-HT_{2C} receptor [61]. These editing events result in codon changes with functional repercussion on transcripts and peptides [7, 35], creating a multitude of protein isoforms, in a fine-tuning mechanism which regulates protein-protein interactions and mRNA expression [10, 56].

Editing has been implicated in development and differentiation [62–65], alternative splicing [66] and stress response, namely hypoxia and viral defense.

Raitskin *et al.* showed that ADARs are complexed with ribonucleoproteins (RNPs) that are involved in the splicing machinery [67]. ADAR2 may interact with the C-terminal domain of RNA polymerase II during transcription (**Figure 14**), which coordinates editing with splicing [66].

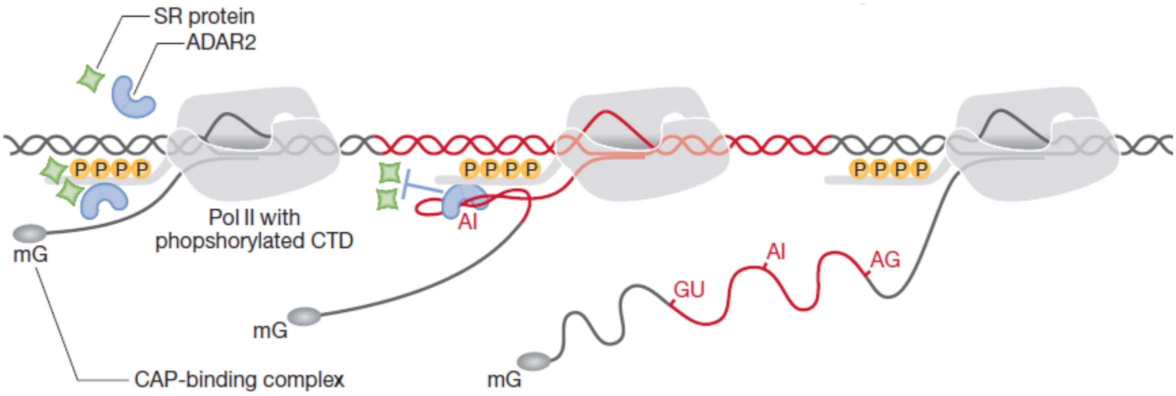


Figure 14. Coordination between editing and splicing. Adapted from [68]. When bound to dsRNA, ADAR2 inhibits binding of serine/arginine-rich (SR) proteins and consequent recruitment of the spliceosome.

Stress conditions such as viral infections and hypoxia lead to upregulation of ADAR1, which is involved in several responses [69].

On one hand, ADAR1 may regulate gene expression through A to I editing, resulting in alternative splicing or nuclear retention of edited transcripts. Hyperedited RNAs are recognised by protein p54^{nrb}, which promotes nuclear retention of inosine-containing transcripts in structures named paraspeckles [70, 71]. These paraspeckles constitute nuclear reservoirs of mRNAs that can be readily exported when required, allowing a faster recovery of normal cellular functions without having to transcribe those genes [72]. Alternatively, hyperedited transcripts may induce the Vigilin complex, which also contains ADAR1. Vigilin associates with a protein responsible for trimethylating histone H3 on lysine 9 and stimulates the formation of heterochromatin (**Figure 15**) [73, 74].

On the other hand, ADAR1 may regulate gene expression independently of RNA editing. It was demonstrated that a region that encompasses the Z-DNA binding domain and the first dsRBD of ADAR1 may be associated with NF90 and control gene expression [75].

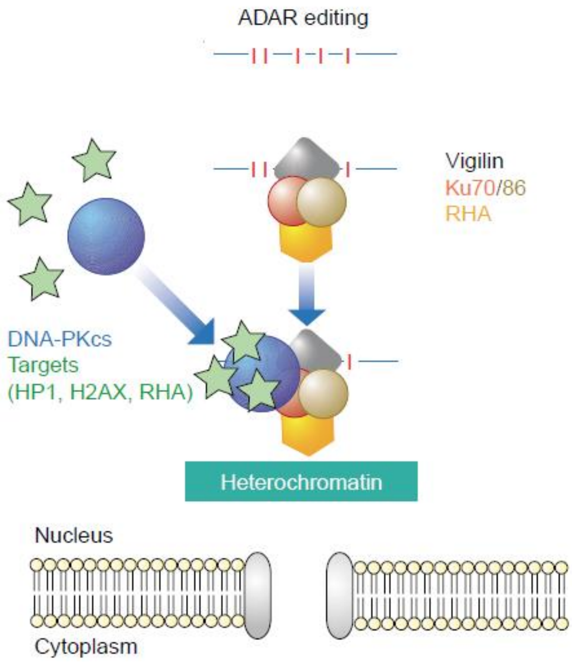


Figure 15. Nuclear retention of hyperedited RNAs. Adapted from [76].

In the presence of viral RNA, editing may result in quite different outcomes.

Viral dsRNA may trigger interferon production, which induces a promoter of the gene that codes ADAR1 and produces the longer isoform ADAR1L. It has been conjectured that Z-DNA-binding domains are required to bind these dsRNAs in order to promote their hyperediting by the deaminase domain [8, 19]. Hyperedited RNAs carrying a higher number of I-U mismatches are efficiently cleaved by the RISC component Tudor SN ribonuclease [77], reducing the probability of these RNAs expressing viral epitopes. Hyperedited viral RNA may activate specific Toll-like receptors, which stimulate an inflammatory response [78]. Furthermore, stress response is induced through activation of PKR kinase, which phosphorylates the translation initiation factor eIF-2α and halts protein translation [79, 80].

Editing may also inhibit protein synthesis required for assembly and release of the virus, causing its persistence instead [12, 19]. However, some viruses have adapted to editing and may exploit its occurrence to proliferate in a host and assemble viral particles [81]. In some cases, mutation of viral coding sequences inhibits viral suppression by PKR kinase and hence inhibits stress response, increasing the host's susceptibility [80].

Antiviral response and gene expression may also be regulated by RNA interference (RNAi). ADAR1 may edit primary forms of microRNA (miRNA) and siRNA, which results in a less efficient cleavage by Drosha to precursor RNAi hairpins [10, 82]. These hairpins may also be edited before Dicer processing, which affects the complementarity and production of mature RNAi molecules (**Figure 16**) [12, 19, 82]. Complementarity is essential for RNAi efficiency and target specificity. Therefore, ADAR1 may create new regulatory targets, however it may also reduce efficiency by competing with the RISC complex for binding to these RNAs [80, 82].

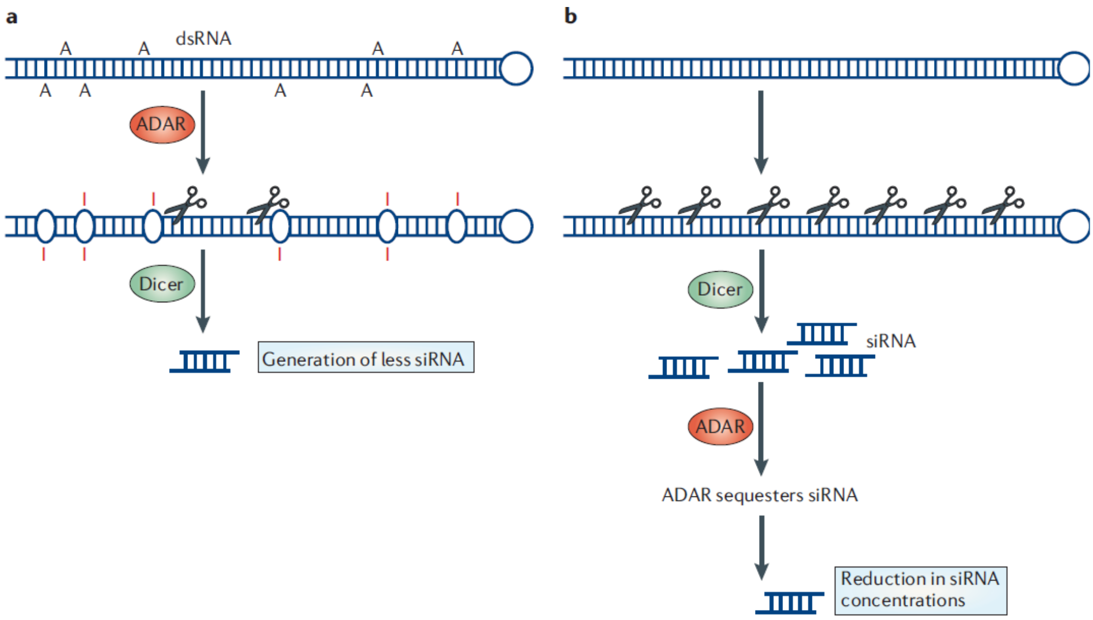


Figure 16. ADAR1 may create mismatches and reduce the length of dsRNA fragments that can be cleaved to siRNA **(a)** or sequester it **(b)**, which reduces its probability to decay the target mRNA [82].

Nevertheless, Ota *et al.* showed that a monomer of ADAR1 may dimerise specifically with Dicer through a region around its second dsRBD, and increase its maximum rate of precursor miRNA cleavage and subsequent miRNA loading into RISC [83]. Therefore, ADAR1 may also promote RNAi, independently of A to I editing since it would be required an ADAR1 homodimer.

ADAR1 has a central role in regulating immune responses, which may elucidate its extreme significance in mammals by promoting correct hematopoiesis and avoiding interferon overproduction [84]. Therefore, RNA editing may optimise cellular functions and biological pathways, in order to increase chances of survival [10].

The myriad of functions of ADARs in complex biological pathways across several species brings us to question how and why ADARs became so essential.

One explanation could reside on the Baldwin effect, proposed by James Mark Baldwin [85] and reintroduced by George Gaylord Simpson [86]. According to this evolutionary theory, the likelihood of an individual to acquire new traits throughout life, either via learning or determined through physical interaction with the environment during development, is beneficial for it to adapt faster to a new environment. This adaptation requires phenotypic plasticity, which is subject to mutation and selection until converging to a state of optimal fitness [87]. RNA editing may provide randomly pre-adapted transcript variants, which confer phenotypic plasticity without the cost of deleterious mutations at the DNA level [88]. The increased number of variants may lead to different coded proteins and alternatively spliced transcripts, with different biological functions [88].

However, mutations may affect stability and result in selection against introduced editing sites. Only selectively advantageous editing sites will thrive and will probably be improved by additional changes throughout evolution, while deleterious editing sites are removed [88, 89]. The low incidence of RNA editing in coding sequences currently reported may reflect selection against events which are most likely deleterious.

Therefore, RNA editing provides genetic variability and may tolerate mutations at the DNA level by increasing functional redundancy of transcript variants, which leads to a greater complexity [88]. A to I RNA editing tends to reverse G to A mutations in the DNA, resembling the ancestral genome [49, 89].

1.2. ALU ELEMENTS

Repetitive elements represent over 55% of all RNAs transcribed from eukaryotic DNA. One of the most representative types are Alu elements, which are repetitive, high-copy number short interspersed nuclear elements (SINEs) with lengths around 300 nucleotides, that have been generated by reverse transcription and transposed into primates' genomes, comprising nowadays about 10% of the human genome (over 1 million copies) [2, 90].

Alu elements contain highly methylated CpG sites (**Figure 17**), especially in somatic tissues, which promote histone condensation and suppress RNA polymerase III activity in order to maintain low free Alu levels, stabilising nucleosome position and organising chromatin [91, 92].

However, under stress conditions such as heat shock or viral infection, Alu RNAs may be overexpressed and inhibit RNA polymerase II, hence repressing general gene transcription [92, 93]. Particularly, antisense Alu elements bind to heat shock factors in upstream regions of genes, repressing protein synthesis [94]. Furthermore, these retroelements can be transcribed and inserted in other sites in the host genome, influencing transcription of host genes by providing alternative promoters, splice sites or stop codons.

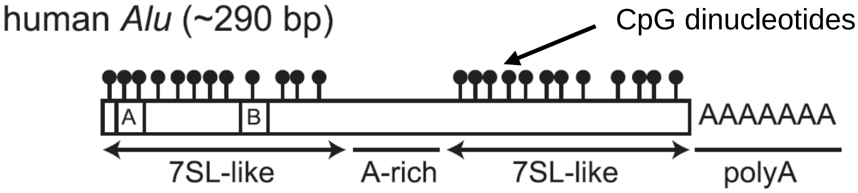


Figure 17. Structure of human Alu elements. Adapted from [91].

Alu elements are generally included in the host genome, either in coding regions, introns or UTRs, with preference for noncoding intronic regions (**Figure 18b**) [95]. Their insertion depends mostly of cell stress to activate transcription factor binding sites for RNA polymerase III, which results in the transcription of free, noncoding RNAs (**Figure 18a**). These free Alu elements may then be incorporated in other regions of the host genome, although they lack the gene for reverse transcriptase [93].

Intronic Alu elements located in antisense orientation may be transcribed and regulate gene transcription of the sense strand or modulate splicing and translation initiation [96]. Moreover, inserted Alu elements in the 3'-UTR may create alternative polyadenylation sites and be targeted by miRNA [97, 98]. Formation of stable secondary structures is also crucial for repression of gene expression through stress granule association, which have associated miRNAs to block translation of Alu-containing mRNAs [99]. In addition, free long noncoding Alu elements may bind to inserted Alu elements, allowing recognition by protein Staufen1, which mediates decay of transcriptionally active Alu-containing mRNAs [100].

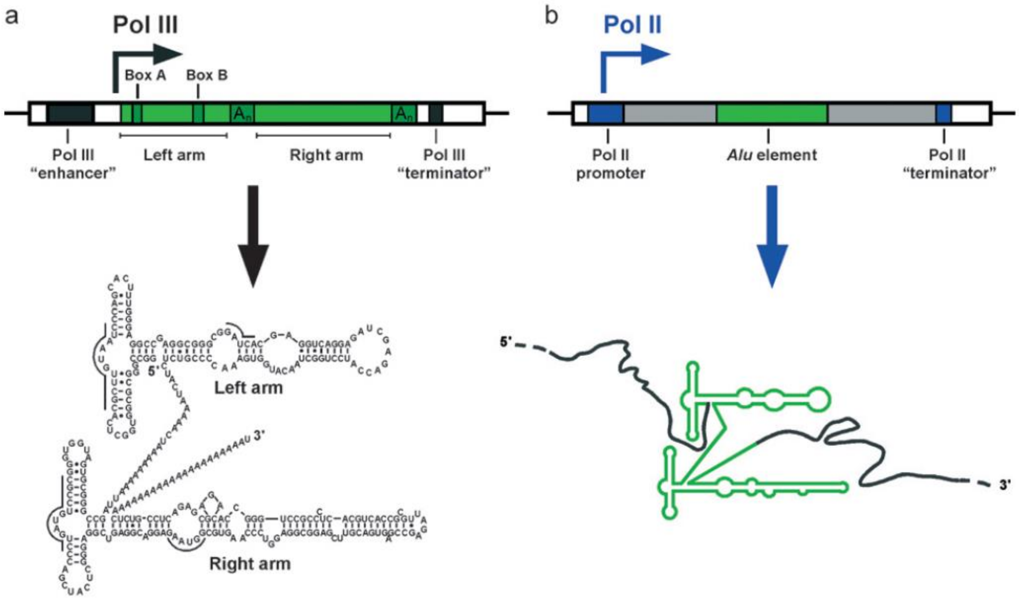


Figure 18. Transcription of Alu elements by RNA polymerases II and III. Alu can be transcribed as free RNAs **(a)** or inserted into the host genome **(b)** [93].

The majority of RNA editing occurs in UTRs and introns, promoted especially by large regular duplexes formed between inverted repeats [19]. The repetitive nature of Alu elements promotes the formation of stable secondary structures and, in association with their preferred location in these genomic regions, implies them as possible candidates to undergo A to I editing (**Figure 19**).

Athanasiadis *et al.* demonstrated that an extensive occurrence of A to I RNA editing in Alu elements throughout the transcriptome, particularly in genes, implying significant effects on cellular gene expression. Furthermore, they showed that editing occurs typically at intramolecular stem loops formed between inverted Alu repeats, whereas editing in secondary structures resulting from interactions between Alu elements from different mRNA molecules was rare [44].

The formation of these structures in an intron may shift the splicing pattern of the downstream exon from constitutive to alternative splicing [101]. Alternatively, editing may create new splice sites in Alu elements, promoting its insertion as an exon, in a process called exonisation.

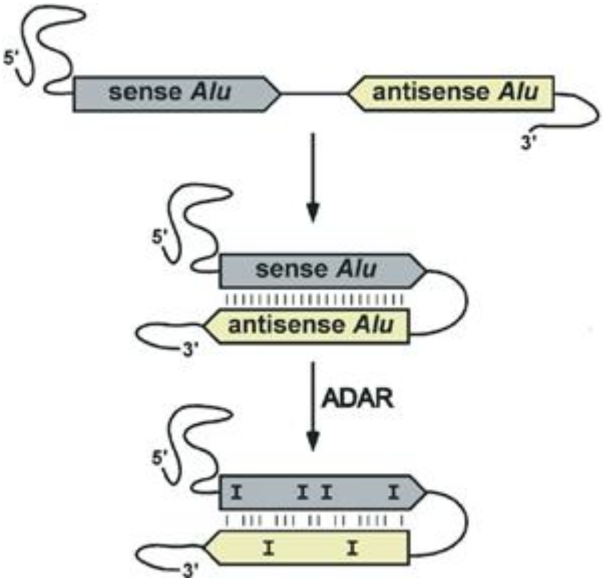


Figure 19. A to I RNA editing of Alu elements [96].

Alternative splicing of Alu exons increases transcriptomic diversity and allows the existence of new isoforms while maintaining the original isoform, which contribute to less selection pressure and may confer evolutionary advantage [102, 103]. Transposition of currently active Alu elements might contribute to the evolutionary future of humans and other primates [104, 105].

However, such evolutionary events are maintained at low inclusion levels, because some aberrant transcripts derived from exonisation may be deleterious and suffer negative selection. Furthermore, antisense Alu elements may activate new splice sites through binding of splicing factor U2AF$_{65}$ to polypyrimidine tracts, such as the poly(U) sequences derived from poly(A) present in sense Alu elements (**Figure 20**) [106]. U2AF$_{65}$ stimulates binding of U2 snRNP to the 3'-SS, promoting the recruitment of the spliceosome complex [107]. Excessive Alu exonisation is avoided through competition between this splicing factor and heterogeneous nuclear RNP (hnRNP) C, which represses exon inclusion [106].

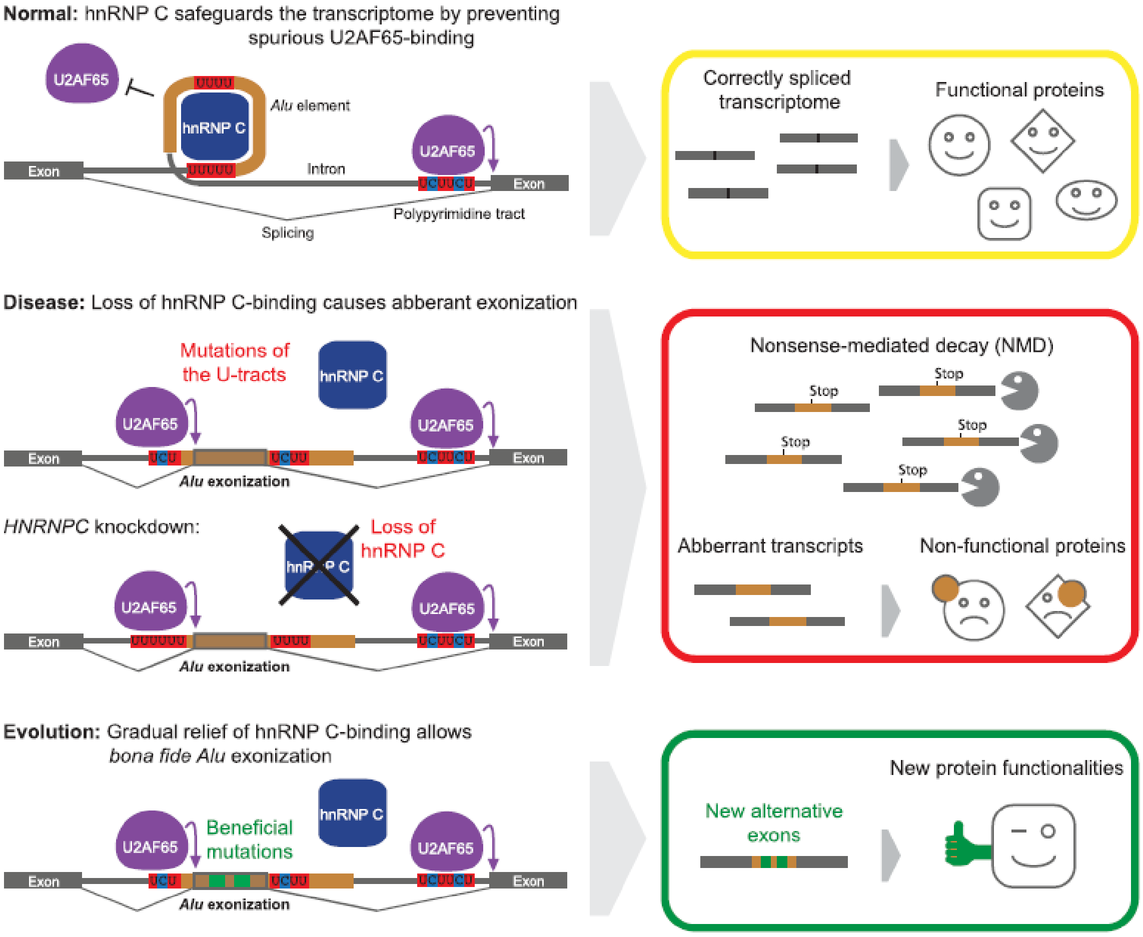


Figure 20. Regulation of Alu element exonisation [106].

RNA editing in inverted Alu pairs may induce site-selective editing in nearby adenosines with a low basal level of editing [108], which may promote modifications in coding sequence and splicing patterns, as mentioned above.

The influence of Alu elements in alternative splicing ignites a special interest in splicing regulation and transcriptome diversity.

1.3. CIRCULAR RNA

Although discovered more than 20 years ago, circular RNA (circRNA) were considered to be rare configurations which resulted probably from splicing errors or artificial transcriptional noise [109], since they did not show any associated biological function, as they lack the 5'-cap and the poly(A) tail required for translation [110]. Nevertheless, recently they have emerged as a common and naturally abundant form of noncoding RNA, conserved in *Eukarya*, *Bacteria* and *Archaea* [111–113].

CircRNAs are stated as new splicing patterns where, instead of the canonical donor GU at the 5'-SS and the acceptor AG at the 3'-SS of a single intron between two exons, some RNAs present the donor GU at the 5'-SS of a downstream intron and the acceptor AG at the 3'-SS of an upstream intron, promoting its circularisation (**Figure 21**) [109, 113, 114]. Circularisation may also result from self-splicing of group I introns, which would then reintegrate into other mRNAs [111], or backsplicing events between both splice sites of the same exon [113].

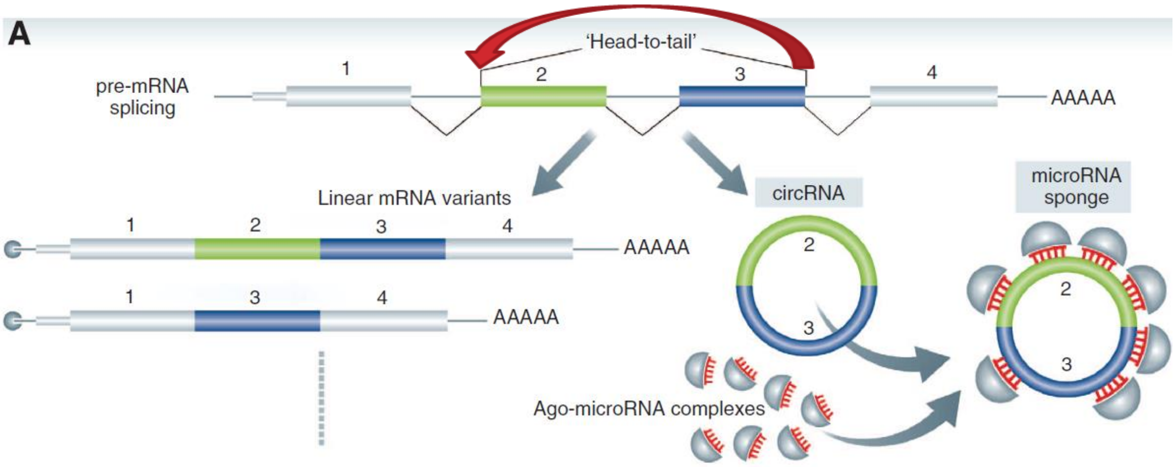


Figure 21. Biogenesis of circRNA from alternative splicing. Adapted from [114].

Memczak *et al.* analysed human transcripts and obtained a set of RNA sequencing reads linked with circRNAs, most of which have the same genomic orientation as known genes, whereas smaller fractions are antisense to known transcripts, or UTRs, introns and some unannotated regions of the genome [112]. It has been suggested that circRNAs probably compete with other RNAs for miRNA binding [112, 115]. Hansen *et al.* showed an example of an antisense circRNA that captures miRNAs that would block antisense transcripts from constituting duplexes with the sense mRNA, suppressing its expression [116].

Moreover, it is suggested that circRNAs may also bind to RNA-binding proteins (RBPs), modulating their free concentration and of their targets [112]. Despite being regarded primarily as noncoding RNAs, Chen and Sarnow demonstrated *in vitro* that an internal ribosome entry site (IRES) inserted in circRNAs enables them to be recognised by the ribosome, resulting in the translation of long repeating polypeptide chains for multiple consecutive rounds (**Figure 22**) [117]. This mechanism of translation may be exploited by viruses inside the host [118].

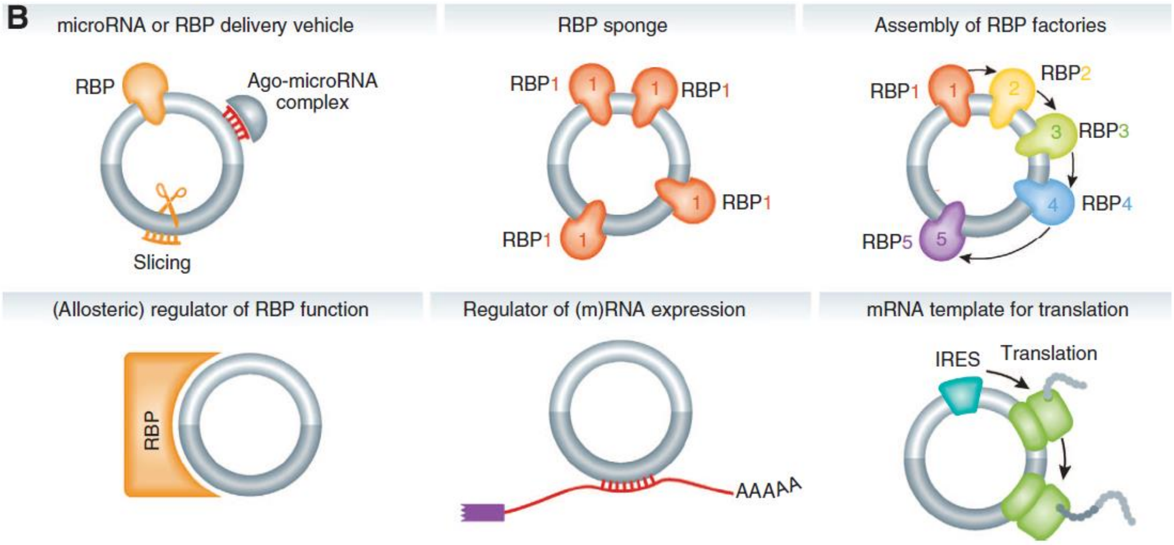

B
microRNA or RBP delivery vehicle
RBP
Ago-microRNA complex
Slicing
RBP sponge
RBP1
1
Assembly of RBP factories
RBP1
RBP2
RBP3
RBP4
RBP5
1
2
3
4
5
(Allosteric) regulator of RBP function
RBP
Regulator of (m)RNA expression
AAAAA
mRNA template for translation
IRES
Translation

Figure 22. Other potential functions of circRNAs [114].

Abundance of circRNAs in the human transcriptome implies that these structures may have important functions, as they appear to be specifically expressed across tissues or developmental phases which present significant enrichment of circRNA sequences with conserved nucleotides [109, 112, 113]. Circular structures tend to be more resistant to degradation from exoribonucleases and therefore may be more expressed in the cytoplasm [109, 113, 115].

Despite circRNAs may be transcribed throughout the genome, Salzman *et al.* noticed that these sequences were more frequently derived from exons [109]. Jeck *et al.* observed that circRNAs are more likely to be flanked by Alu repeats, which led them to postulate that base pairing between complementary inverted Alu repeats in long flanking introns may promote RNA circularisation [113]. Indeed, complementarity between Alu elements may contribute to RNA folding back, bringing both splice sites closer (**Figure 23**).

1.4. AIMS OF THIS WORK

Regarding all the information presented, it is appealing to conjecture that Alu elements may have significant influence on alternative splicing and formation of circRNA. Therefore, we postulate that A to I editing in secondary structures formed by inverted Alu repeats potentially influences circularisation.

With their extensive repertoire of potential biological functions, ADARs may have significant regulatory functions, not only in expressing several protein isoforms and controlling their expression, but also in regulating alternative splicing and synthesising regulatory circRNAs, which may influence RNAi.

The involvement of both Alu elements and ADAR in alternative splicing suggests their potential role on circRNA formation.

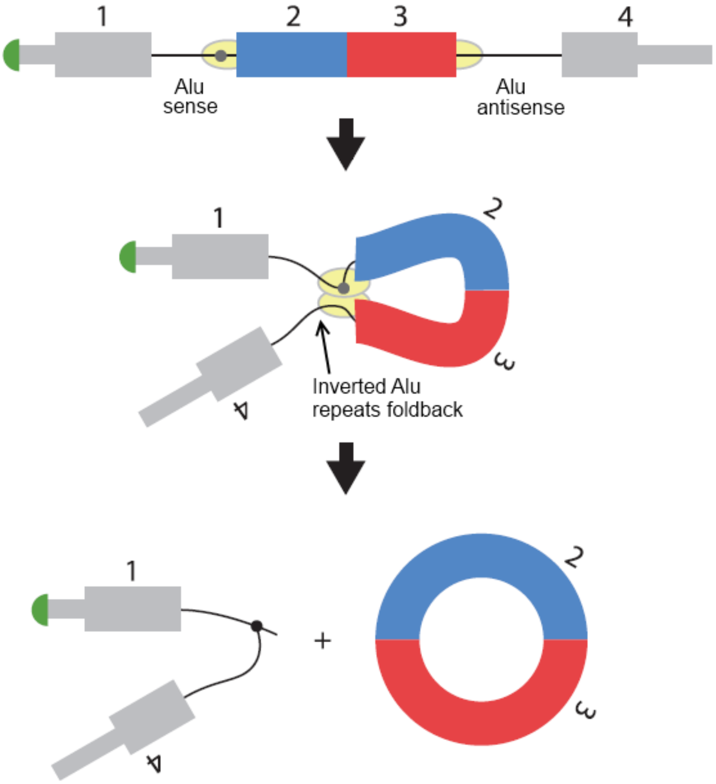


Figure 23. Possible model of RNA circularisation. Adapted from [106].

This work consists on a computational analysis of these potential circRNAs obtained in several studies, in order to evaluate the influence of A to I editing in inverted Alu pairs flanking these RNAs.

2. MATERIALS AND METHODS

2.1. TOOLS AND COMPUTATIONAL RESOURCES

For our analysis, we used the following software packages on UNIX-based systems:

- Python 2.7.6
- NumPy 1.8.1 + SciPy 0.13.3
- Biopython 1.63 + EMBOSS 6.6.0
- R 3.0.2 + RPy2 2.3.9

Python is a high-level programming language, with a strong amount of abstraction from the syntax of the machine language, which makes it very clear and easy to read. Furthermore, its versatility and open source license allows it to be widely used for scripting purposes [119]. We designed a graphical user interface (GUI) to manage data parsing, processing and analysis, which were all implemented in Python, supplemented with several packages (**Appendix A, Figure A1**).

NumPy and SciPy are collections of packages for numerical computation in mathematics, science and engineering [120]. NumPy and SciPy were used to treat numerical data and perform statistical analysis, in combination with R through the RPy2 package [121].

R is an open source language and environment which is widely used for statistical computing and graphics, providing a myriad of powerful statistical functions in the built-in packages [122].

Biopython is a package written in Python which is designed for biological computation [123]. Biopython was applied to parse and treat sequence information from FASTA files regarding Alu elements, in order to determine potential inverted pairs through their complementarity. For this purpose, we executed the Smith-Waterman alignment algorithm [124] from the EMBOSS package [125], in order to obtain the best local alignment between inverted Alu elements.

2.2. OBTAINING AND PARSING DATA

We designed Python scripts in order to obtain and parse data from Memczak *et al.* [112] and Jeck *et al.*, with the same stringency cutoffs as described [113]. We obtained 4 datasets comprising genomic coordinates of the annotated circRNAs: Memczak, Jeck (*Low*), Jeck (*Medium*) and Jeck (*High*).

We obtained the full set of annotated genes in the human genome from the knownGene track in the UCSC Genome Browser (GRCh37/hg19 assembly, February 2009) [126]. This set comprised not only the genomic coordinates of each gene, but also the number and coordinates of each annotated exon. We identified several gene isoforms through the kgXref cross-reference table. For genes with multiple isoforms, we chose the isoform with the most exons (**Figure 24**). We only considered genes from fully annotated somatic chromosomes (chr1-chr22) and sex chromosomes (chrX, chrY). Therefore, we obtained a set of 28,842 unique gene isoforms.

We extracted the coordinates of exons inside these unique gene isoforms and determined intron coordinates between the end of an upstream exon and the start of a downstream exon. We obtained 233,456 exons and 205,742 introns (**Appendix A, Figure A2**).

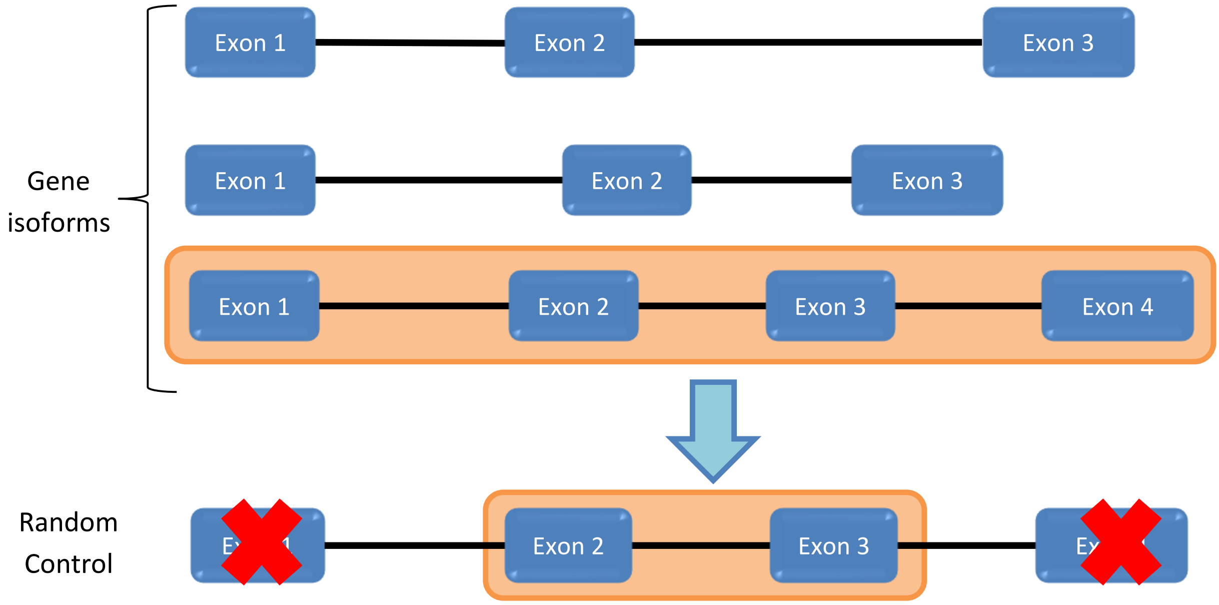


Figure 24. Computational pipeline for the creation of reference exons and introns, and control.

We defined our control as a random selection of possible combinations of one exon or multiple exons separated by introns (**Figure 24**). In order to become potential circular forms, we assume that both start and end coordinates should match splice sites. Therefore, we obtained our control from a set of 183,622 internal exons, which was filtered from the exon set by excluding the first and the last exon of each gene isoform (**Appendix A, Figure A2**).

2.3. ANALYSIS OF CIRCRNA ARCHITECTURE AND GENE ONTOLOGY

Using the set of exons and introns, we determined and compared the gene architecture of datasets provided by both groups with our controls (**Appendix A, Figure A2**). We considered that circRNAs contain specific exons and introns if their genomic coordinates are between the start and end circRNA coordinates. If exons and introns are only partially inside a circRNA, their length is truncated from the circRNA boundary in order to determine exonic and intronic content, respectively (**Figure 25**).

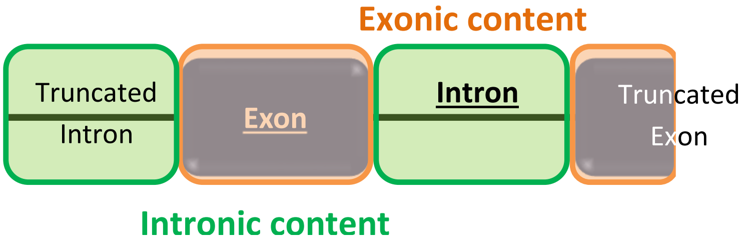


Figure 25. Characterisation of content of a circRNA. Full exons and introns are highlighted in bold and underlined. Exonic content is highlighted in orange and intronic content in green.

Because genes may share the genomic coordinates of some exons and/or introns, if a circRNA contained exons from multiple genes, we determined to which gene the circRNA belonged by demanding the following criteria: (1) splice site coincidence (2, 1 or 0 ends); (2) highest number of exons between circRNA boundaries. The latter was used only if the number of splice site matches was the same for multiple genes, from which we considered only the first gene in the list if there were tied genes after the application of both criteria. After determining to which gene each circRNA belongs, it was possible to determine which introns were included in that circRNA.

Because these datasets had different variances, we performed the non-parametric Kruskal-Wallis test [127] for each evaluated parameter, with *post-hoc* Wilcoxon-Mann-

Whitney U tests [128] with continuity correction (**Figures 29, 31-34**). Significance levels were corrected for multiple hypotheses testing using the Bonferroni correction [129]:

$$P_B = \frac{P}{m} \quad ; \quad m = \frac{n \times (n-1)}{2}$$

where P_B is the Bonferroni-corrected p-value, P is the obtained p-value, m is the number of comparisons and n is the number of datasets which were analysed. In plots containing error bars, these represent the respective standard error of the mean.

Based on the differences among datasets, we filtered and defined a dataset of *Proper* circRNAs, whose start and end genomic coordinates match exon boundaries, similarly to our control. This dataset was used for further analyses (**Appendix A, Figure A2**).

We determined gene ontology of *Proper* circRNAs using the Database for Annotation, Visualisation and Integrated Discovery (DAVID)'s Gene Functional Classification tool [130].

2.4. ANALYSIS OF ALU ELEMENTS FLANKING CIRCRNAS

We obtained a set of 1,175,329 Alu elements in the human genome, from the RepeatMasker track of the UCSC Genome Browser [131]. We considered only 613,563 Alu elements which were located in the gene isoforms filtered previously (**Appendix A, Figure A3**). We searched for Alu elements in both upstream and downstream introns of circRNAs, if applicable (**Figure 26**).

We defined 30 random control datasets with the same sample size as our circRNA dataset, and compared the abundance of Alu elements in introns flanking circRNAs. The number of controls was chosen based on the computational effort required to provide the analyses with a high statistical power, which represents the probability of reporting a significant difference (P < 0.05) as significant [132].

We considered a flanking distance – under which Alu elements would be completely included – from 500 bases to 5 kb for each side, with 500-base intervals (**Appendix A, Figure A3**).

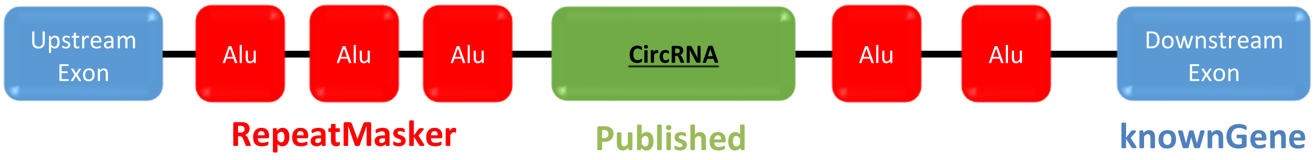


Figure 26. Illustration of the analysis of the flanking regions of a provided circRNA.

In order to define which Alu elements on each flank would be available to form an inverted pair with an Alu element of the other flank, we analysed their orientation and filtered out those which could form inverted pairs within the same flank (**Figure 27**). We simulated the possible inverted Alu pairs within the same flank by determining to which Alu element with the opposite orientation each Alu element would pair, following these criteria: (1) genomic distance between Alu elements; (2) complementarity between Alu elements. The latter was used to break ties only – such as tandem Alu elements – and we considered that a higher similarity score in the Smith-Waterman alignment would confer a greater stability to the inverted pair, thus improving its likelihood. If 2 competing alignments had the same similarity score, we would compare how many nucleotides of the alignment would be complementary (% identity).

We considered circRNAs with at least 1 potential inverted Alu pair if they have available at least 1 plus (+) and 1 minus (-) stranded Alu element on opposite flanks (**Figure 27**).

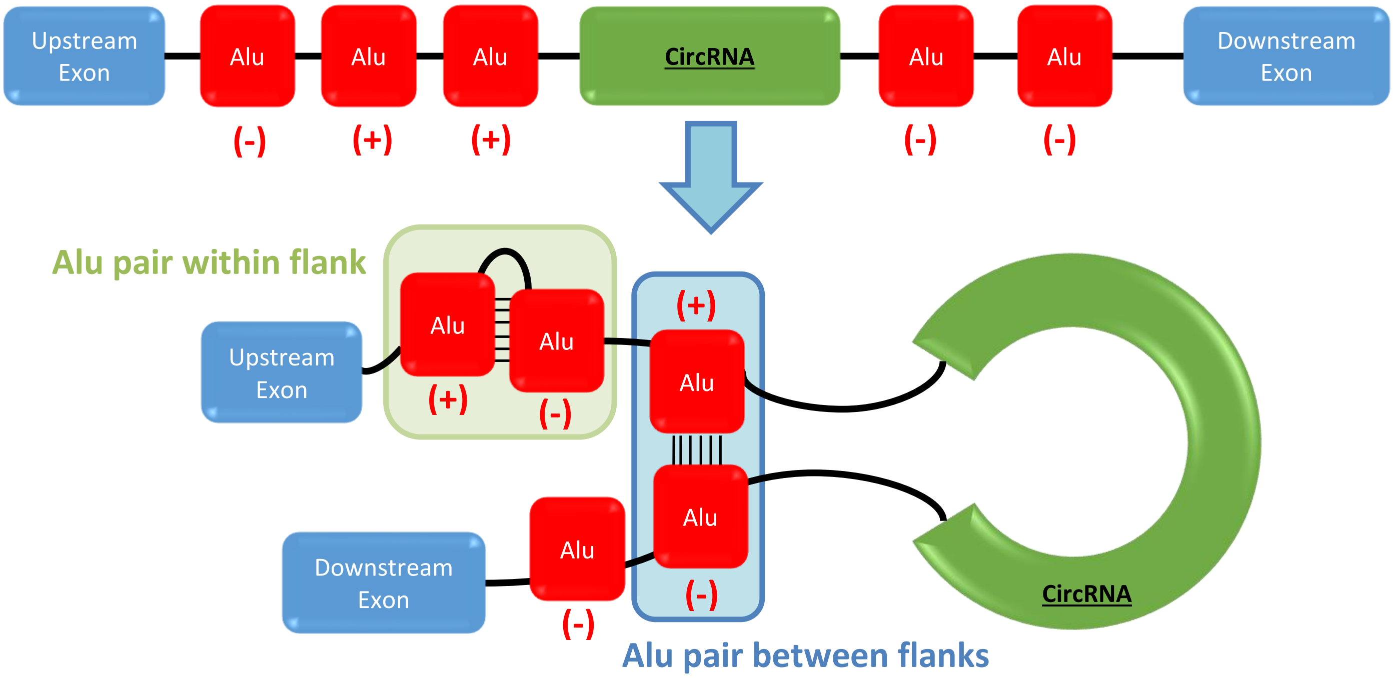


Figure 27. Illustration of a circRNA with inverted Alu pairs.

We also simulated which inverted Alu pairs would be likely to form between flanks, but only considering genomic distance to solve possible pairs. We performed the Smith-Waterman local alignment and obtained the correspondent similarity score, which would indicate the stability of the inverted Alu pairs formed between flanks, considering the possible presence of indels and mismatches (**Appendix A, Figure A4**).

Then, we explored the incidence of A to I RNA editing in circRNAs which have at least 1 inverted Alu pair, and identified which Alu elements are edited (**Appendix A, Figure A5**).

For that purpose, we used PREFA, which is a software platform developed by Athanasiadis *et al.* in order to obtain a reference set of edited Alu elements throughout the human transcriptome [44]. This software platform searches for clusters of A to G mismatches between cDNA obtained from GenBank mRNAs and RefSeq DNA sequences (**Figure 28**), which result from RNA editing involving secondary structures in the same mRNA.

These mismatches could represent SNPs or sequencing errors in databases, which required a stringent approach in order to reduce the likelihood of false positives by only selecting clusters of at least 5 A to G transitions, in the absence of other base discrepancies [44]. Located mismatches were subjected to a χ^2 test comparing the observed number of A to G transitions with the expected probability of an A to G discrepancy to occur. If the test statistic was higher than the critical value derived from a significance level α, the observed A to G mismatches were considered to result from editing [44].

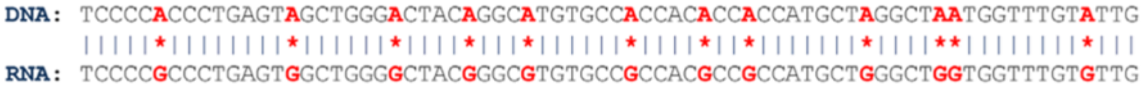

DNA: TCCCCACCCTGAGTAGCTGGGACTACAGGCATGTGCCACCACACCACCATGCTAGGCTAATGGTTTGTATTG
|||||*||||||||*||||||*||||*|||*||||||*||||*||*|||||||*||||**||||||||*|||
RNA: TCCCCGCCCTGAGTGGCTGGGGCTACGGGCGTGTGCCGCCACGCCGCCATGCTGGGCTGGTGGTTTGTGTTG

Figure 28. Identification of mismatches between RNA (cDNA) and DNA [133].

We obtained a set of 3,298 edited Alu elements using a cutoff where the critical value was obtained for $\alpha = 10^{-5}$, which represents a probability of 1 in 100,000 observed A to G mismatches not resulting from RNA editing.

We compared the genomic coordinates of Alu elements flanking circRNAs with at least 1 inverted pair with the editing set, in order to detect which Alu elements were edited with a high significance level (**Appendix A, Figure A5**).

Furthermore, we analysed the orientation of Alu elements flanking circRNAs with at least 1 inverted pair, in order to assess whether there was a bias for either sense or antisense Alu elements, which would minimise the formation of inverted pairs within the same flank and rather promote inverted pairs between flanks (**Appendix A, Figure A6**). Therefore, we analysed each flank of a circRNA and calculated the difference between the number of Alu elements on one orientation and the number of Alu elements on the other, and compared with the orientation of the circRNA to verify if there was a trend for Alu elements to be either sense or antisense. We considered 1 kb distance bins in order to minimise potential bias due to the reduced distance window where Alu elements could be located.

Statistical significance of our comparisons was analysed by 2 tests:

- One-sample t-test [134] between the mean of 30 controls and the value obtained in the circRNA dataset, after confirming the normality assumption for more than 90% of the distributions of control values with the Shapiro-Wilk test [135], which is the most powerful normality test [136] (**Figures 35a, 36, 37, 39a, 45**);
- Non-parametric Wilcoxon-Mann-Whitney U test for unknown distributions of both circRNAs and control (**Figures 35b, 38, 39b, 40, 41, 43, 44**).

In all plots, error bars represent the respective standard error of the mean.

3. Results and Discussion

3.1. CircRNA architecture and dataset quality control

We compared datasets of circRNAs provided by Memczak *et al.* [112] and Jeck *et al.* [113] with our control.

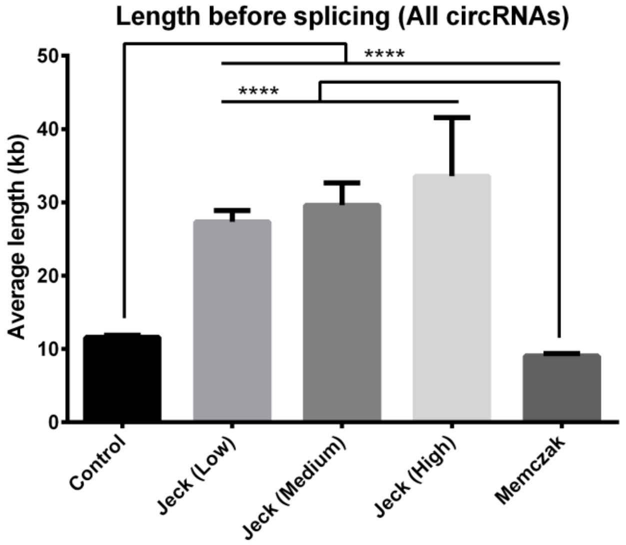


Figure 29. Length of all circRNAs before splicing. (****), $P < 10^{-4}$.

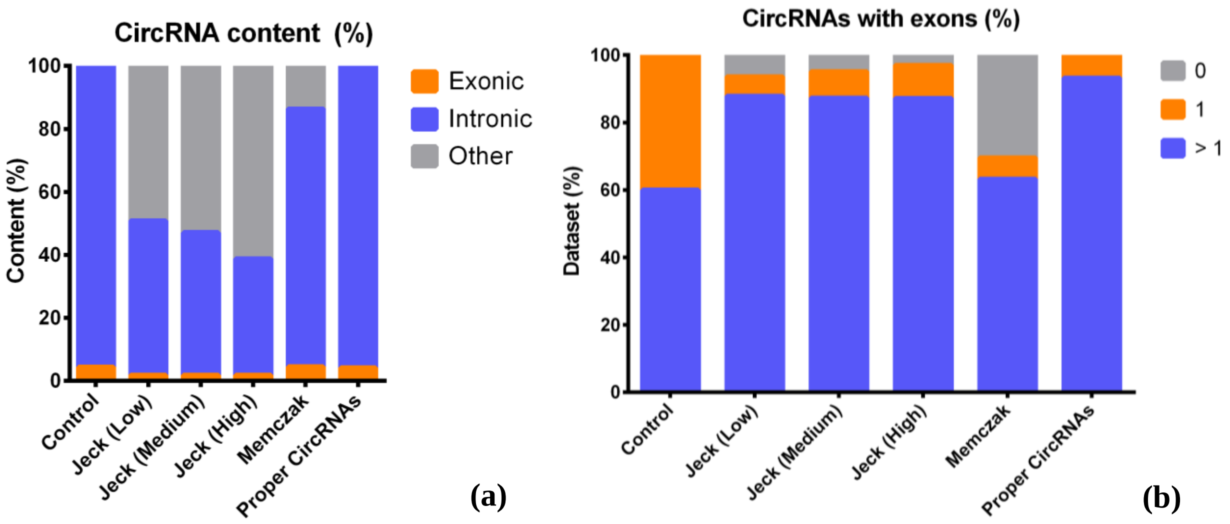


Figure 30. Presence of aberrant circRNAs in datasets from both papers.

We observed highly significant differences in average length before splicing among datasets, implying discrepancies in the detection of circRNAs (**Figure 29**).

Those discrepancies were more accentuated when evaluating the nature of these circRNAs. Significant portions of these circRNAs were not identified as part of exons or introns (**Figure 30a**). More than 50% of circRNA content in the datasets provided by Jeck *et al.* and approximately 10% of circRNA content in the dataset provided by Memczak *et al.* are unannotated regions or regions regarded as intergenic. Moreover, significant portions of circRNAs from both datasets do not contain exons (**Figure 30b**).

We cannot exclude the possibility that some of the identified circRNAs which do not contain exons are, in fact, circular intronic long noncoding RNAs (ciRNAs). Zhang *et al.* described these forms as lariat-derived introns that escaped from debranching after splicing, which localise to the nucleus [137]. Therefore, formation of ciRNAs appears to be independent of the formation of circRNAs, which result from exon backsplicing. This led us to exclude circRNAs which did not contain any exon, despite their potential regulatory functions in gene expression [137].

Nevertheless, the observed unknown content may be explained by errors occurring in RNA sequencing and respective mapping of the reads. Both groups performed mapping of paired-end reads after generating cDNA libraries. This approach enables the acquisition of sequence information from two points in a transcript with an estimated distance between the reads. However, reverse transcription is not subject to a strict regulation as the biological forward transcription, which may result in cDNA products which contain fragments from both strands and may therefore not align properly to the reference sequence, leading to false positives [138].

Jeck *et al.* performed the library sequencing using longer reads, which allow better mapping and alignment to the reference sequence [113]. These reads were mapped using an independent algorithm with high sensibility and sensitivity in the detection of novel splice sites (MapSplice) [139]. This might have reduced the probability of erroneous mapping and detection of some truncated forms with partial exons. However, read mapping is subject to uncertainty due to sequencing errors and other technical artifacts, which create ambiguity in the alignment [138].

On the other hand, Memczak *et al.* created their own methodology and applied specific criteria, such as excluding abnormally large circRNAs (> 100 kb), which explains the lower average length compared to Jeck *et al.* (**Figure 29**) [112].

These differences among datasets led us to compare which circRNAs were common to datasets provided by both groups, and to create a subset containing only circRNAs whose start and end genomic coordinates match exon boundaries (*Proper* circRNAs).

Table 1. Comparison between datasets from Jeck *et al.* and Memczak *et al.*

Dataset	CircRNAs	Proper	Common	Overlapping	
Jeck (Low)	7769	6184 (79.60%)	549* (7.07%)	1584 (13.32%)	Memczak
Jeck (Medium)	2228	1828 (82.05%)	331 (14.86%)	628 (13.33%)	
Jeck (High)	485	420 (86.60%)	159 (32.78%)	214 (11.34%)	
Memczak	1951	1071 (54.89%)	549* (28.14%)	878 (16.86%)	Jeck (Low)
			331 (16.97%)	563 (11.89%)	Jeck (Medium)
			159 (8.15%)	241 (4.20%)	Jeck (High)

* 492 of 549 circRNAs (89.62%) were identified as *Proper* circRNAs.

Small fractions of these datasets had the same genomic coordinates and orientation (**Table 1**). More than 80% of circRNAs provided by Jeck *et al.* were identified as *Proper*, whereas only about 55% of circRNAs provided by Memczak *et al.* passed the criterion, which is consistent with published results [112].

Approximately 90% of the circRNAs which were found in datasets provided by both groups were also identified as *Proper* circRNAs, which reflects the quality of our criteria. Based on the high absolute and relative frequencies of circRNAs that pass this condition (**Table 1**), we decided to use *Proper* circRNAs provided by Jeck *et al.* at a *Low* cutoff [113] for further analyses.

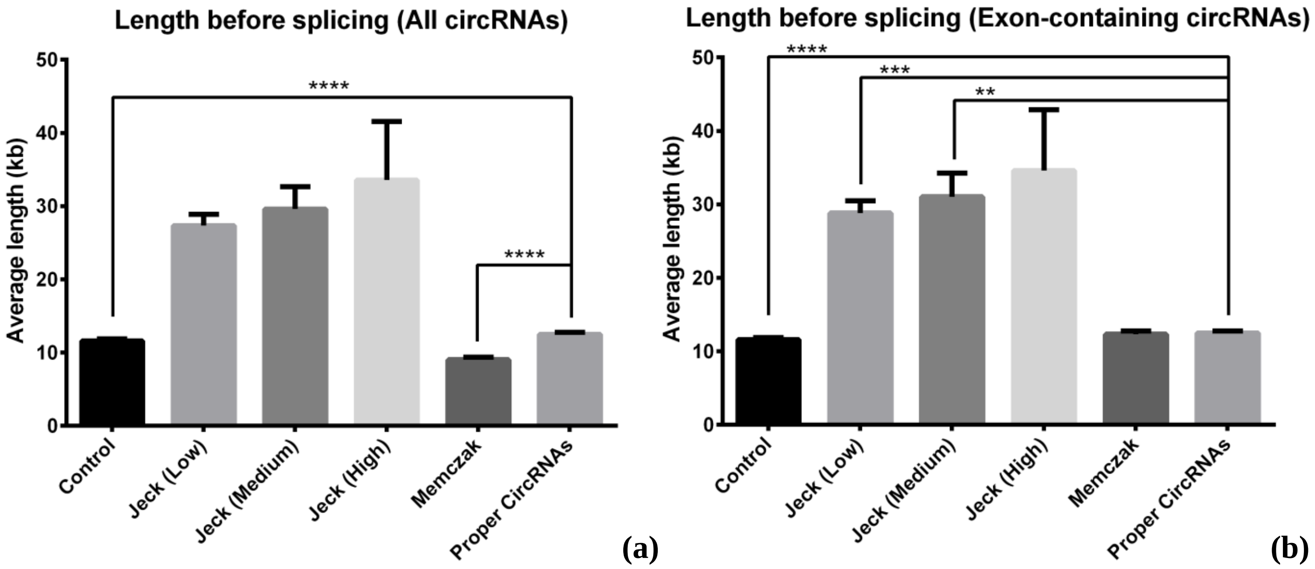


Figure 31. Length of all circRNAs **(a)** and circRNAs which contain exons **(b)** before splicing. P-values relatively to *Proper* circRNAs: (****), $P < 10^{-4}$; (***), $P < 10^{-3}$; (**), $P < 10^{-2}$. Control is significantly different from all other datasets (****, $P < 10^{-4}$).

Creating a filter for *Proper* circRNAs resulted in values closer to the control. Unspliced *Proper* circRNAs have significantly different lengths from the respective not filtered dataset (**Figure 31**). CircRNAs provided by Memczak *et al.* have similar length to *Proper* circRNAs and control, when considering only circRNAs which have full exons (**Figure 31b**). When considering all circRNAs, they are significantly shorter than *Proper* circRNAs and control (**Figure 31a**), which implies the abundance of short truncated exons with splice sites different from annotated, especially when only about 55% of these circRNAs were identified as *Proper* circRNAs (**Table 1**).

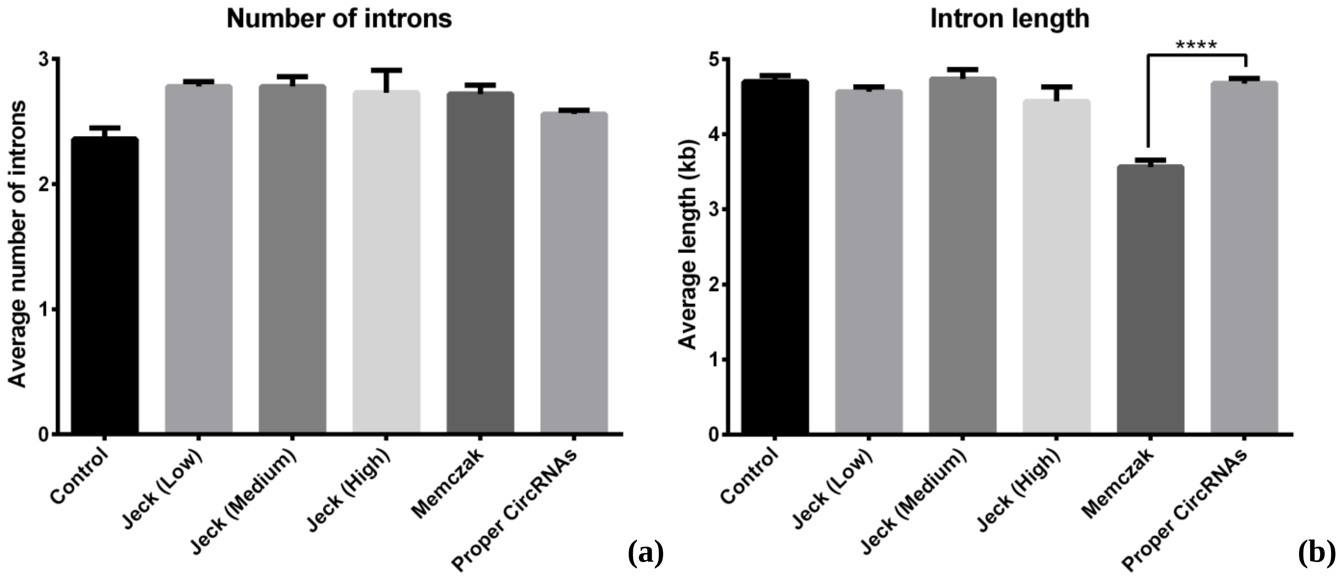


Figure 32. Number **(a)** and length **(b)** of introns in circRNAs which contain exons. Only in **(b)** there were significant differences among datasets. P-value relatively to *Proper* circRNAs: (****), $P < 10^{-4}$.

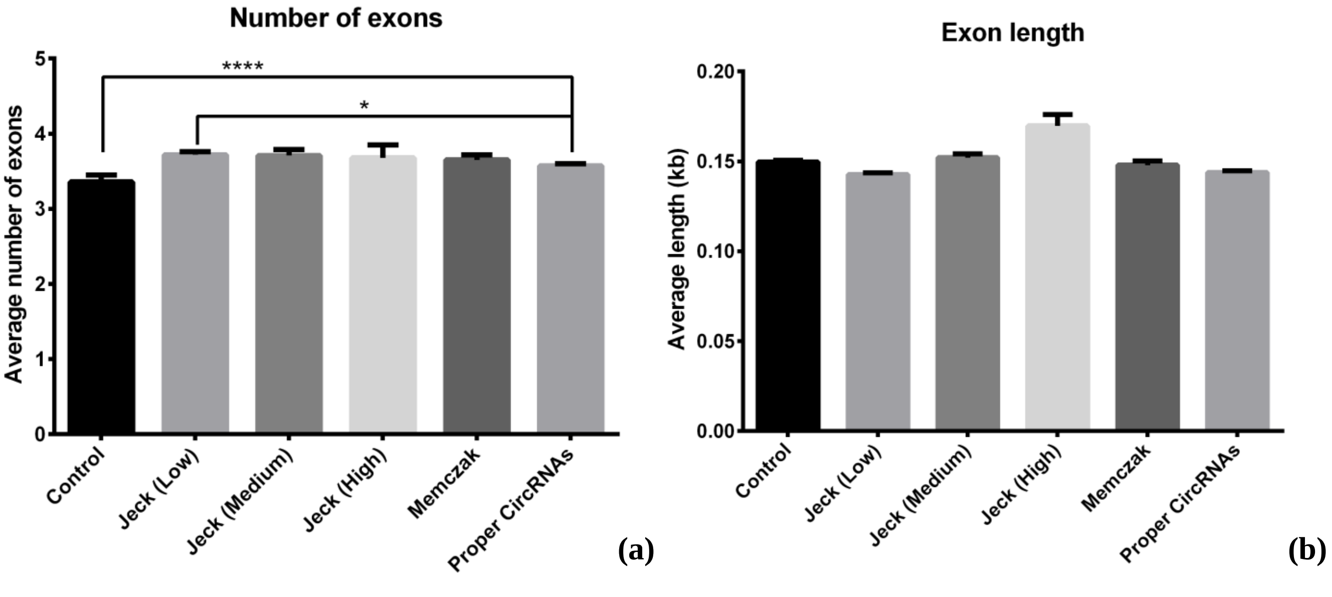


Figure 33. Number **(a)** and length **(b)** of exons in circRNAs which contain exons. Plotted P-values: (****), $P < 10^{-4}$; (*), $P < 0.05$. Control is significantly different from all other datasets in **(a)** (****, $P < 10^{-4}$).

Significantly small differences between *Proper* circRNAs and control regarding number and length of introns (**Figure 32**) and number of exons (**Figure 33a**), together with the insignificant difference regarding the length of exons (**Figure 33b**), suggest that *Proper* circRNAs have similar gene architecture to randomly chosen internal exons.

Memczak *et al.* circRNAs have significantly shorter introns than other datasets, which may again imply the abundance of short truncated forms with splice sites different from annotated.

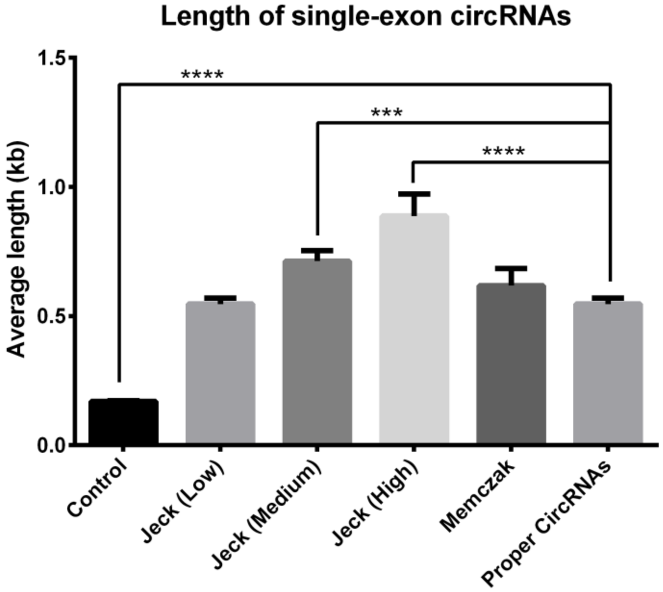


Figure 34. Average length of circularised exons. P-values relatively to *Proper* circRNAs: (****), $P < 10^{-4}$; (***), $P < 10^{-3}$. Control is significantly different from all other datasets (****, $P < 10^{-4}$).

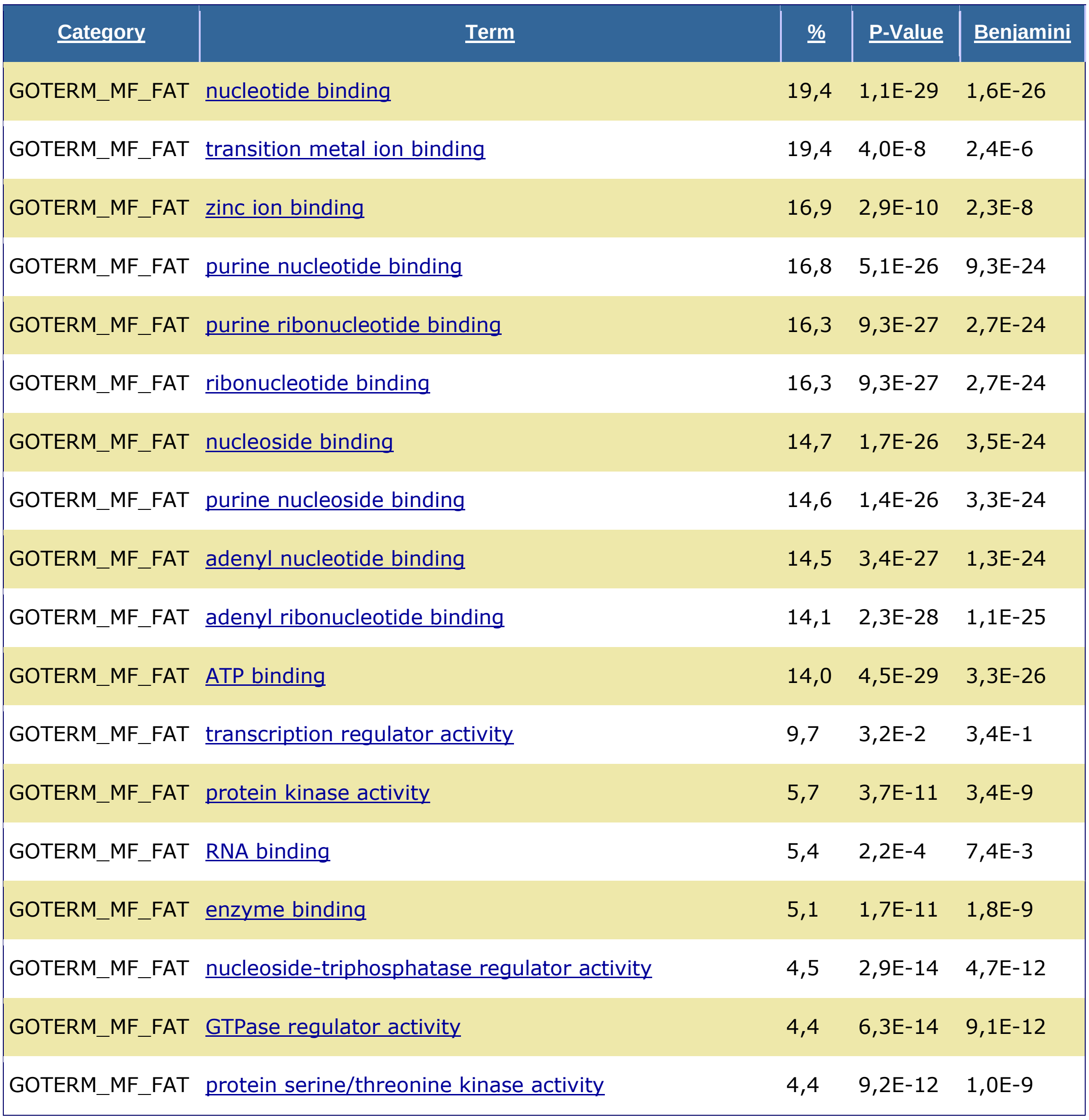
CircRNAs have similar patterns of gene architecture to randomly chosen internal exons.

Only in the case of single-exon circRNAs, we observe significantly large differences between control and all circRNA datasets (**Figure 34**). *Proper* single-exon circRNAs are approximately 3-fold longer than control, which is consistent with published data [113]. Their length may be explained by their location, as *Proper* circRNAs have more propensity to be located as upstream as possible, starting preferably in the second exon of a gene (**Appendix B, Table B1**) [109]. Large internal exons tend to be either the second or the penultimate exon, which appear to derive from terminal exons that are much longer than internal exons [140], and could influence the formation of local secondary structures that regulate alternative splicing [141]. Considering evaluated parameters, circRNA biogenesis is probably linked to their vicinity, namely the presence of Alu elements as reported [113].

3.2. Gene ontology

We determined gene ontology of circRNAs using DAVID's Gene Functional Annotation tool [130]. We observed that 1,861 of 3,131 genes (59.44%) were highly enriched for encoding phosphoproteins (P-value = 1.4×10^{-210}, Benjamini corrected P-value = 9.3×10^{-208}), which may act more significantly as protein kinases. Therefore, our *Proper* circRNAs maintained the enrichment for genes with kinase activity and nucleotide binding (**Table 2**), observed by Jeck *et al.* [113].

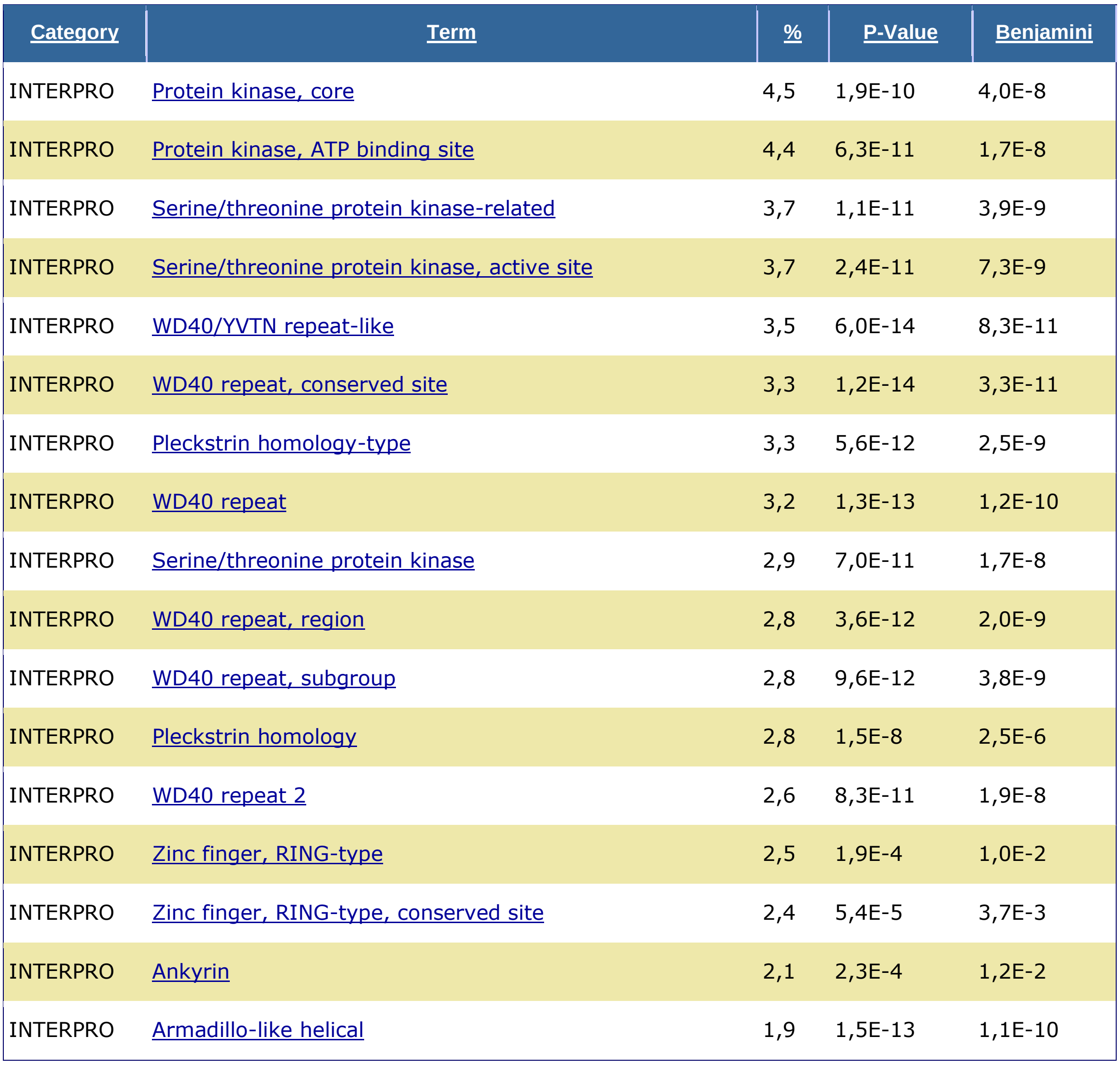
Table 2. Some of the most representative molecular functions of genes which contain circRNAs.



Category	Term	%	P-Value	Benjamini
GOTERM_MF_FAT	nucleotide binding	19,4	1,1E-29	1,6E-26
GOTERM_MF_FAT	transition metal ion binding	19,4	4,0E-8	2,4E-6
GOTERM_MF_FAT	zinc ion binding	16,9	2,9E-10	2,3E-8
GOTERM_MF_FAT	purine nucleotide binding	16,8	5,1E-26	9,3E-24
GOTERM_MF_FAT	purine ribonucleotide binding	16,3	9,3E-27	2,7E-24
GOTERM_MF_FAT	ribonucleotide binding	16,3	9,3E-27	2,7E-24
GOTERM_MF_FAT	nucleoside binding	14,7	1,7E-26	3,5E-24
GOTERM_MF_FAT	purine nucleoside binding	14,6	1,4E-26	3,3E-24
GOTERM_MF_FAT	adenyl nucleotide binding	14,5	3,4E-27	1,3E-24
GOTERM_MF_FAT	adenyl ribonucleotide binding	14,1	2,3E-28	1,1E-25
GOTERM_MF_FAT	ATP binding	14,0	4,5E-29	3,3E-26
GOTERM_MF_FAT	transcription regulator activity	9,7	3,2E-2	3,4E-1
GOTERM_MF_FAT	protein kinase activity	5,7	3,7E-11	3,4E-9
GOTERM_MF_FAT	RNA binding	5,4	2,2E-4	7,4E-3
GOTERM_MF_FAT	enzyme binding	5,1	1,7E-11	1,8E-9
GOTERM_MF_FAT	nucleoside-triphosphatase regulator activity	4,5	2,9E-14	4,7E-12
GOTERM_MF_FAT	GTPase regulator activity	4,4	6,3E-14	9,1E-12
GOTERM_MF_FAT	protein serine/threonine kinase activity	4,4	9,2E-12	1,0E-9

Furthermore, we determined the main domains of proteins encoded by genes that contain circRNAs (**Table 3**). We observed an enrichment for domains characteristic of protein kinases, observed in Table 2, and other significant binding domains.

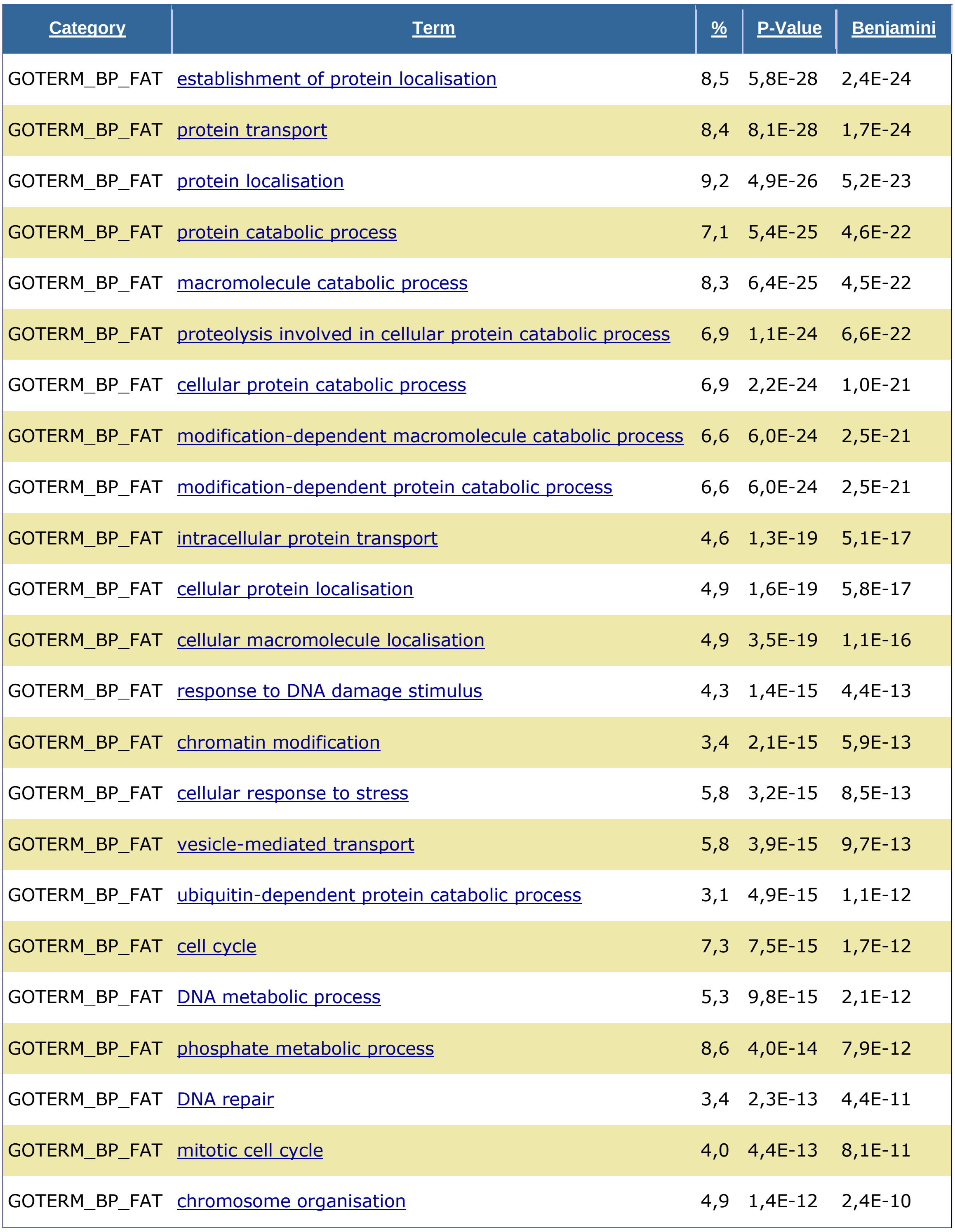
Table 3. Main protein domains encoded by genes which contain circRNAs, according to the InterPro database [142].



Category	Term	%	P-Value	Benjamini
INTERPRO	Protein kinase, core	4,5	1,9E-10	4,0E-8
INTERPRO	Protein kinase, ATP binding site	4,4	6,3E-11	1,7E-8
INTERPRO	Serine/threonine protein kinase-related	3,7	1,1E-11	3,9E-9
INTERPRO	Serine/threonine protein kinase, active site	3,7	2,4E-11	7,3E-9
INTERPRO	WD40/YVTN repeat-like	3,5	6,0E-14	8,3E-11
INTERPRO	WD40 repeat, conserved site	3,3	1,2E-14	3,3E-11
INTERPRO	Pleckstrin homology-type	3,3	5,6E-12	2,5E-9
INTERPRO	WD40 repeat	3,2	1,3E-13	1,2E-10
INTERPRO	Serine/threonine protein kinase	2,9	7,0E-11	1,7E-8
INTERPRO	WD40 repeat, region	2,8	3,6E-12	2,0E-9
INTERPRO	WD40 repeat, subgroup	2,8	9,6E-12	3,8E-9
INTERPRO	Pleckstrin homology	2,8	1,5E-8	2,5E-6
INTERPRO	WD40 repeat 2	2,6	8,3E-11	1,9E-8
INTERPRO	Zinc finger, RING-type	2,5	1,9E-4	1,0E-2
INTERPRO	Zinc finger, RING-type, conserved site	2,4	5,4E-5	3,7E-3
INTERPRO	Ankyrin	2,1	2,3E-4	1,2E-2
INTERPRO	Armadillo-like helical	1,9	1,5E-13	1,1E-10

WD40 repeat, Pleckstrin homology, Ankyrin, Armadillo and zinc finger domains are among those which explain some of the major biological functions (**Table 4**) and pathways (**Table 5**) calculated by DAVID. These domains are responsible for a variety of cellular processes, such as the regulation of transcription and mRNA trafficking, cellular compartmentalisation of some proteins, cell adhesion and signal transduction, and cell cycle regulation, either through the organisation of the cytoskeleton, either through chromatin remodeling and protein folding. All these biological functions are highlighted in Table 4.

Table 4. Main established functions of genes which contain circRNAs.

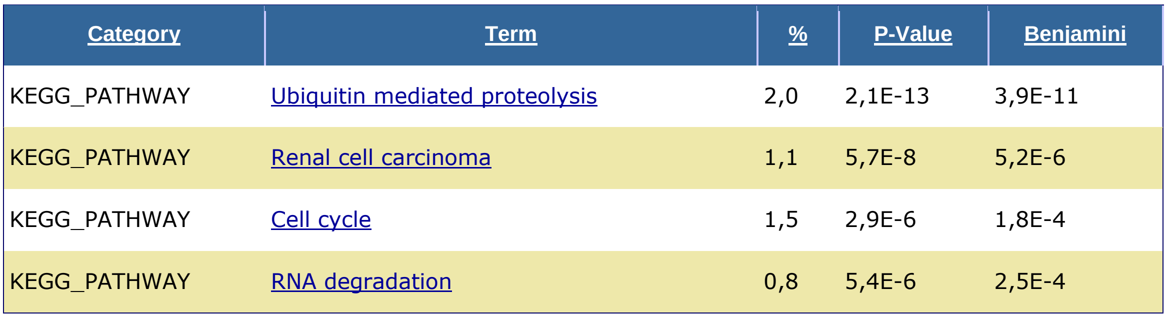


Category	Term	%	P-Value	Benjamini
GOTERM_BP_FAT	establishment of protein localisation	8,5	5,8E-28	2,4E-24
GOTERM_BP_FAT	protein transport	8,4	8,1E-28	1,7E-24
GOTERM_BP_FAT	protein localisation	9,2	4,9E-26	5,2E-23
GOTERM_BP_FAT	protein catabolic process	7,1	5,4E-25	4,6E-22
GOTERM_BP_FAT	macromolecule catabolic process	8,3	6,4E-25	4,5E-22
GOTERM_BP_FAT	proteolysis involved in cellular protein catabolic process	6,9	1,1E-24	6,6E-22
GOTERM_BP_FAT	cellular protein catabolic process	6,9	2,2E-24	1,0E-21
GOTERM_BP_FAT	modification-dependent macromolecule catabolic process	6,6	6,0E-24	2,5E-21
GOTERM_BP_FAT	modification-dependent protein catabolic process	6,6	6,0E-24	2,5E-21
GOTERM_BP_FAT	intracellular protein transport	4,6	1,3E-19	5,1E-17
GOTERM_BP_FAT	cellular protein localisation	4,9	1,6E-19	5,8E-17
GOTERM_BP_FAT	cellular macromolecule localisation	4,9	3,5E-19	1,1E-16
GOTERM_BP_FAT	response to DNA damage stimulus	4,3	1,4E-15	4,4E-13
GOTERM_BP_FAT	chromatin modification	3,4	2,1E-15	5,9E-13
GOTERM_BP_FAT	cellular response to stress	5,8	3,2E-15	8,5E-13
GOTERM_BP_FAT	vesicle-mediated transport	5,8	3,9E-15	9,7E-13
GOTERM_BP_FAT	ubiquitin-dependent protein catabolic process	3,1	4,9E-15	1,1E-12
GOTERM_BP_FAT	cell cycle	7,3	7,5E-15	1,7E-12
GOTERM_BP_FAT	DNA metabolic process	5,3	9,8E-15	2,1E-12
GOTERM_BP_FAT	phosphate metabolic process	8,6	4,0E-14	7,9E-12
GOTERM_BP_FAT	DNA repair	3,4	2,3E-13	4,4E-11
GOTERM_BP_FAT	mitotic cell cycle	4,0	4,4E-13	8,1E-11
GOTERM_BP_FAT	chromosome organisation	4,9	1,4E-12	2,4E-10

A previous study showed that some of these identified biological functions and proteins were most frequently influenced in alternative splicing, suggesting these genes may be subject to fine adjustments of the resultant protein functions [143].

Furthermore, Shen *et al.* observed a preferential influence of Alu elements in alternative splicing of upstream exons in zinc finger genes, which have essential functions in regulating transcription [105]. Considering that in the genomic context circRNAs start preferably at exon 2 [109], Alu elements in the 5'-UTR may influence decisively alternative splicing and consequent protein translation of those genes.

Table 5. Major metabolic and regulatory pathways associated with genes which contain circRNAs, according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [144].



Category	Term	%	P-Value	Benjamini
KEGG_PATHWAY	Ubiquitin mediated proteolysis	2,0	2,1E-13	3,9E-11
KEGG_PATHWAY	Renal cell carcinoma	1,1	5,7E-8	5,2E-6
KEGG_PATHWAY	Cell cycle	1,5	2,9E-6	1,8E-4
KEGG_PATHWAY	RNA degradation	0,8	5,4E-6	2,5E-4

Therefore, circRNAs may represent essential forms which are required to control important regulatory pathways in the cell. Among the most representative metabolic pathways enriched in genes which contain circRNAs are the ubiquitin mediated proteolysis and RNA degradation pathways (**Table 5**).

The ubiquitin mediated proteolysis pathway is a protein quality control system against aberrant nascent polypeptides and misfolded proteins [145]. Most of the genes associated with this pathway are related with the E2 and E3 enzymes, which mediate the direct transfer of ubiquitin onto the target protein [146]. Then, the proteasome bounds to the ubiquitin chain and degrades the target. This pathway may also arise as result of cellular stress which caused protein misfolding and damage, and is essential to regulate protein repertoire to respond to what causes stress and control cell cycle [145].

Genes which contain circRNAs are also enriched in RNA degradation pathways, responsible for quality control of produced transcripts. Some of the main genes which contain circRNAs are involved in the surveillance and decay of aberrantly spliced RNAs, such as those related with the TRAMP complex in exosomes [147], the CCR4-NOT complex [148], the exonuclease Xrn2 [149] and the DDX6 from the decapping complex [150], which result in the turnover of the necessary functional RNAs.

Both of these pathways are essential to regulate cell cycle, especially under stress conditions. During nitrogen starvation in *Schizosaccharomyces pombe*, it was observed that expression of some circRNAs is maintained relatively stable even when their linear mRNA forms were degraded [151]. Therefore, circRNAs may have a conserved function in the regulation of essential genes related with transcription and translation regulation. In addition, the enrichment of circRNA genes for protein kinases, which are regulators of cell proliferation and differentiation, suggests their importance in fundamental regulatory pathways.

3.3. Abundance and editing of Alu elements flanking circRNAs

We compared our circRNAs with the average of 30 random controls with the same sample size and screened for the presence of Alu elements flanking within several genomic distances.

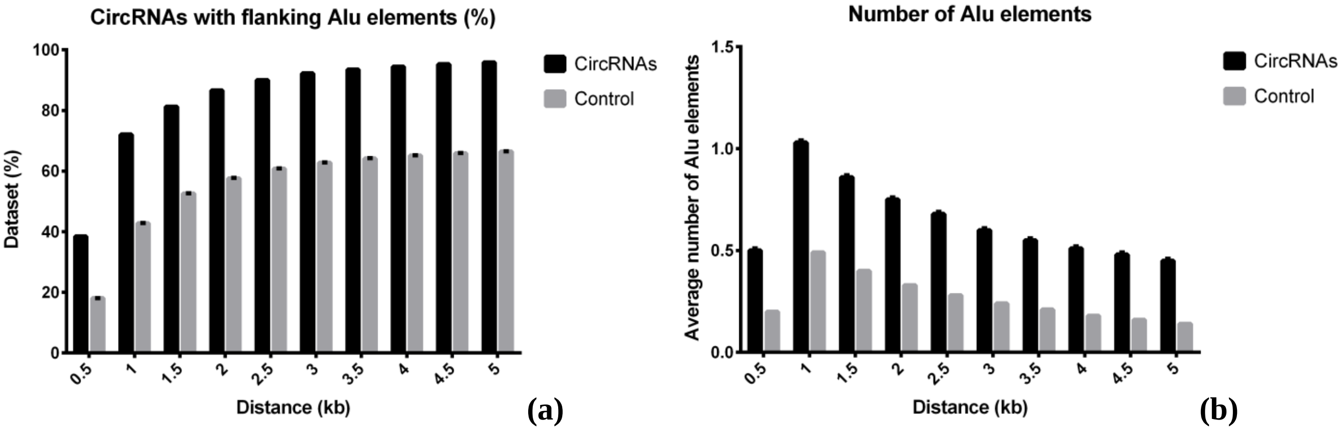


Figure 35. Cumulative number of circRNAs which are flanked by Alu elements **(a)** and average number of Alu elements flanking circRNAs **(b)** by flanking distance. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

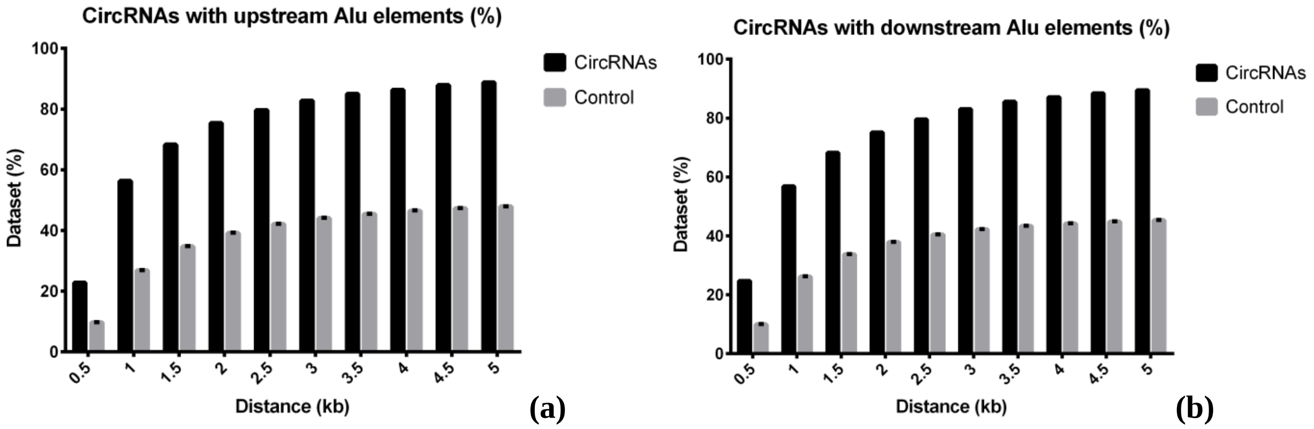


Figure 36. Cumulative number of circRNAs which are flanked by upstream **(a)** and downstream Alu elements **(b)** by flanking distance. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

We observed highly significant differences between circRNAs and control regarding the abundance (**Figure 35a**) and frequency of Alu elements flanking circRNAs (**Figure 35b**), either in upstream (**Figure 36a**) or downstream regions (**Figure 36b**).

In addition, circRNAs are significantly more susceptible to have Alu elements in both flanks (**Figure 37a**), of which a significant portion has Alu elements which are free on one flank to form an inverted pair with Alu elements of the other flank (**Figure 37b**).

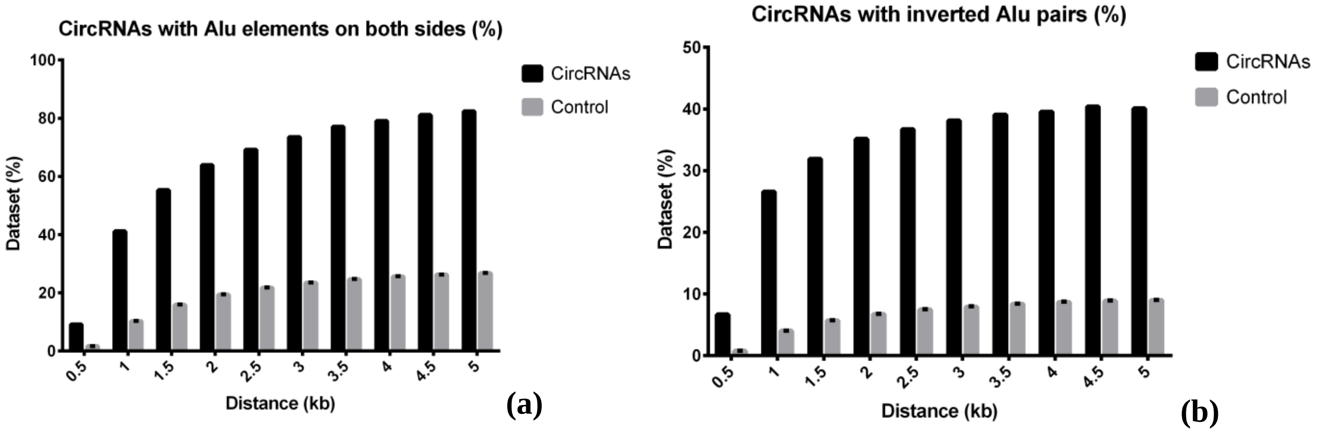


Figure 37. Cumulative number of circRNAs which have Alu elements on both flanks **(a)** and circRNAs which have available Alu elements to form at least 1 inverted pair between flanks **(b)**. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

CircRNAs have significantly more Alu elements in their flanking regions (**Figures 35b, 38a**), which are able to establish more inverted Alu pairs, except in close proximity to the circRNA boundaries (**Figure 38b**). Therefore, we confirmed the reported enrichment for circRNAs to have more Alu elements in their flanking regions [113]. This enrichment is the basis for our hypothesis that stable inverted Alu pairs which may contribute to circularisation might constitute a potential target for A to I RNA editing.

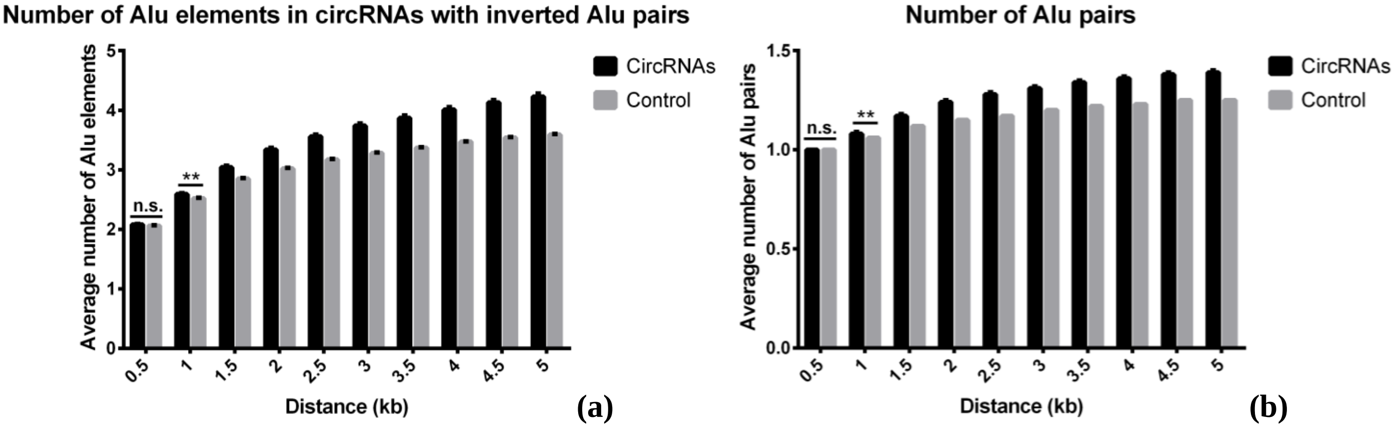


Figure 38. Cumulative average number of available Alu elements in circRNAs which have at least 1 Alu pair **(a)** and predicted number of formed Alu pairs **(b)**. In both plots, plotted p-values: (**), $P < 0.01$; (*n.s.*), $P > 0.05$. Unplotted p-values represent highly significant differences (****, $P < 10^{-4}$).

We used PREFA, which is a software platform developed by Athanasiadis *et al.* in order to obtain a set of edited Alu elements throughout the human genome [44]. We used this set to confirm which Alu elements that are available to form inverted pairs can be edited.

CircRNAs are significantly more predisposed to have at least 1 edited Alu element (**Figure 39a**), which could be explained by the increased probability derived from having significantly more Alu elements which form inverted pairs (**Figure 38**). Incidence of editing may be understood as the editing rate, which stands as:

$$\frac{Number\ of\ edited\ Alu\ elements}{Total\ number\ of\ Alu\ elements} \times 100\%$$

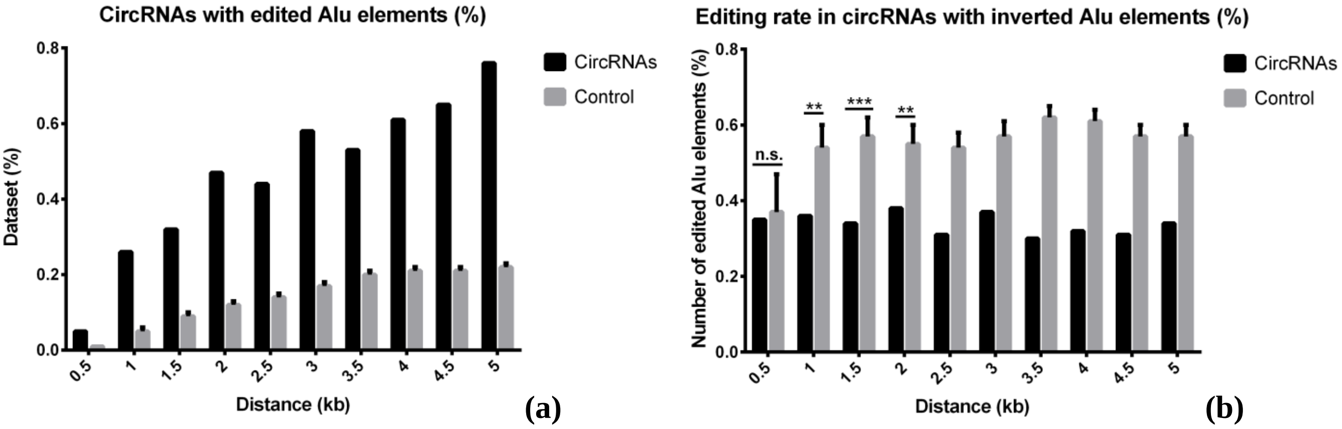


Figure 39. Cumulative number of circRNAs which have edited Alu elements **(a)** and average editing rate in Alu elements flanking circRNAs **(b)**. In both plots, unplotted p-values represent highly significant differences (****, $P < 10^{-4}$). In **(b)**, plotted p-values: (***), $P < 0.001$; (**), $P < 0.01$.

We observe that editing rate in Alu elements flanking circRNAs is significantly lower than in our control, except at close distances (**Figure 39b**).

Moreover, we calculated how heavily these Alu elements were edited:

$$\frac{Number\ of\ editing\ sites}{Length\ of\ edited\ Alu\ element} \times 100\%$$

This ratio indicates editing specificity in an inversely proportional relation: an increased percentage indicates less specific editing, which is frequent in long and stable secondary structures (promiscuous hyperediting); a lower percentage indicates that editing is more specific, probably due to the instability of the secondary structure with bulges and mismatches (site-specific editing).

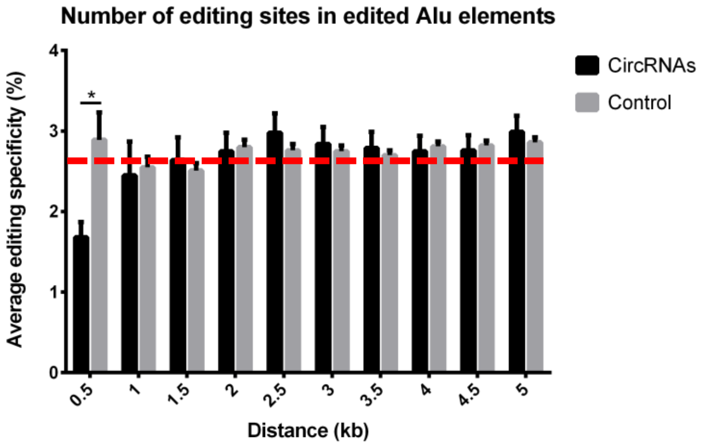


Figure 40. Average editing specificity in edited Alu elements. Plotted p-values: (**), *P* < 0.01; (*), *P* < 0.05. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05). Dashed horizontal line represent a possible plateau basal level of editing specificity.

Due to the repetitive nature of Alu elements, we expected these pairs to be highly stable forms for hyperediting. However, we did not observe significant differences between circRNAs and our controls, except in close proximity to the splice sites, where editing is significantly more selective in circRNAs (**Figure 40**), which either implies that the duplexes formed are less stable or ADARs suffer more constraints in their activity around circRNAs.

We simulated the probable pairs formed between Alu elements from both flanks and calculated their complementarity using the Smith-Waterman local alignment algorithm [124], in order to determine the stability of the Alu pair (**Figure 41**).

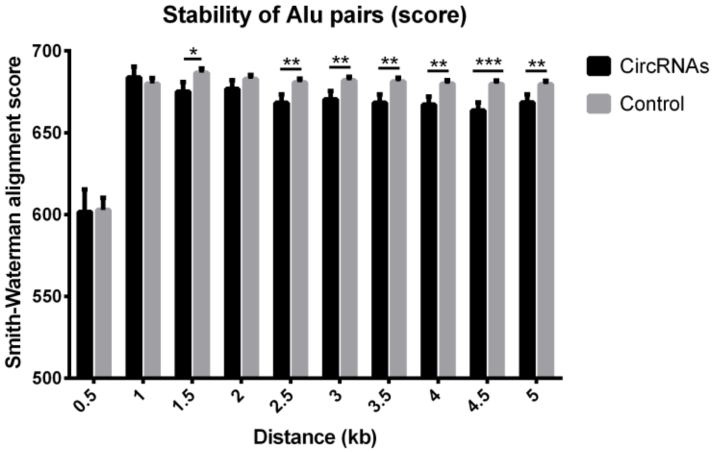


Figure 41. Average stability of Alu pairs. Plotted p-values: (***), *P* < 0.001; (**), *P* < 0.01; (*), *P* < 0.05. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05).

Despite the small differences, these pairs are significantly less stable at longer distances (**Figure 41**). However, if we consider Alu pairs formed within 1 kb away from the respective circRNA, there are no significant differences between circRNAs and our control. Particularly within 500 bp, we observed significant differences in the number of editing sites (**Figure 40**), which cannot be explained by the stability of the inverted Alu pair. The significant reduction of complementarity in both circRNAs and control from 1 kb to 500 bp should reflect a greater editing specificity, which is only observed in circRNAs, while in control these Alu elements are more heavily edited (**Figure 40**). Considering that the average editing selectivity in control is maintained at the same level for all considered distances, the increased selectivity in circRNAs suggests that there are constraints around circRNAs that affect ADAR activity.

ADARs have been associated with spliceosomal components and may perform site-selective editing during pre-mRNA processing [67], coordinated co-transcriptionally by RNA polymerase II [66]. A variety of proteins involved in splicing were shown to have different activities on ADARs. Screens for editing activity of ADAR2 detected DSS1 as a potential interaction platform with proteins such as hnRNPs and stimulate editing [152], and RNA helicase DDX15 as a repressor of RNA editing through unwinding of dsRNA during spliceosome disassembly [153]. Competition between ADARs and spliceosomal components may explain the increased selectivity and lower editing rate in the circRNA dataset (**Figures 39b, 40**). These spliceosomal components would protect these inverted Alu pairs from editing and promote circularisation (**Figure 42**).

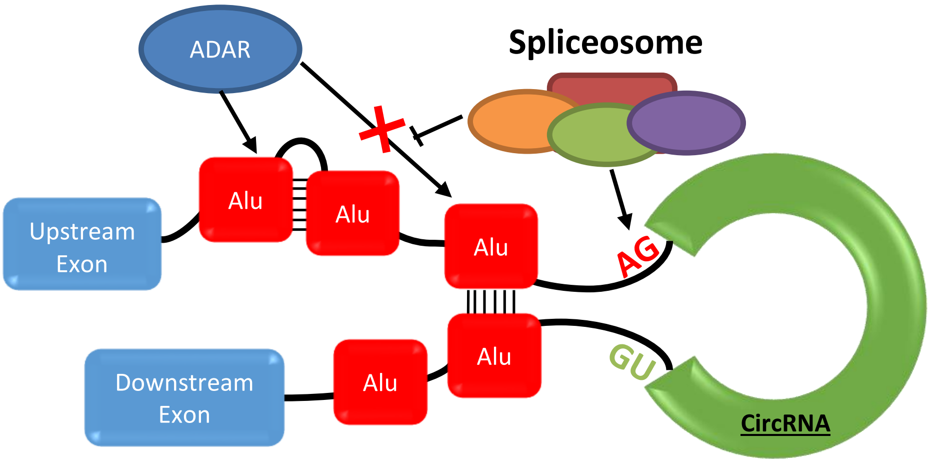


Figure 42. Competition between ADARs and spliceosome may regulate circRNA formation.

On the other hand, secondary structure may have an important role on RNA editing activity. Solomon *et al.* investigated the possible influence of ADARs and Alu elements in alternative splicing and showed that, although editing sites are enriched in alternatively spliced cassette exons over constitutive exons – and respectively in their flanking regions, which contain Alu elements –, ADARs rarely act on essential structures that are required for splicing, such as the branch point nearby the polypyrimidine tract [154]. Editing and the presence of ADARs could affect the stability of secondary structures, thus modifying the availability of splicing regulatory elements and splice sites to splicing factors [154].

RNA structure is very dynamic and sensitive to its surrounding environment. Different conformations may regulate its interaction with other components during gene expression [155]. Wan *et al.* showed that modifications in secondary structures near splice sites may significantly change splicing pattern [156]. Secondary structures that hinder the spliceosome assembly around the splice sites can repress splicing, whereas secondary structures that conceal splicing repressors or bring splicing regulators in close proximity may promote splicing [157, 158].

Secondary structures in upstream regions, where circRNAs are located, are also essential for gene regulation at the 5’-UTR, depending on their stability and position to the 5’-cap in order to bind necessary proteins for translation [159]. The presence of Alu elements flanking circRNAs may also promote the formation of secondary structures near the 5’-UTR.

Secondary structures may be subject to a tight regulation in upstream regions and especially near splice sites, which would hamper ADARs from editing inverted Alu pairs which may be essential to circularisation. On the other hand, the increased editing selectivity in circRNAs – while control sequences appear to maintain the same average number of editing sites in edited Alu elements within all evaluated flanking distances (**Figure 40**) – suggest a potential role of ADARs in solving secondary structures in order not to generate new circRNAs.

It has also been suggested that splicing may be affected by transcription, especially by RNA polymerase II elongation rate and chromatin structure [160]. Veloso *et al.* showed that exon density, GC content and the presence of repeat sequences reduce the speed of RNA polymerase II, suggesting its role in exon definition [161]. Slow elongation rates promote both inclusion or skipping of alternative exons [160, 162], depending on the interactions between sequence motifs and splicing regulators. The accumulation of Alu elements near circularised exons could result in a slow elongation rate, allowing different interactions that lead to different RNA conformations, particularly the looping out of exons that are flanked by Alu elements which could create inverted pairs with Alu elements from different introns.

Furthermore, the presence of antisense Alu elements could represent a greater competition of splicing factors to polypyrimidine tracts [106], resulting in different splicing patterns that could influence circRNA biogenesis. We determined if Alu elements flanking circRNAs have a preferred orientation, which would minimise the formation of inverted pairs within each flanking intron and rather promote pairing between introns.

Therefore, we considered for each flank of a circRNA:

$$Bias = Number\ of\ sense\ Alu\ elements - Number\ of\ antisense\ Alu\ elements$$

We first calculated and compared the averages of the absolute bias for each flank. The absolute bias is the absolute value of the difference mentioned above.

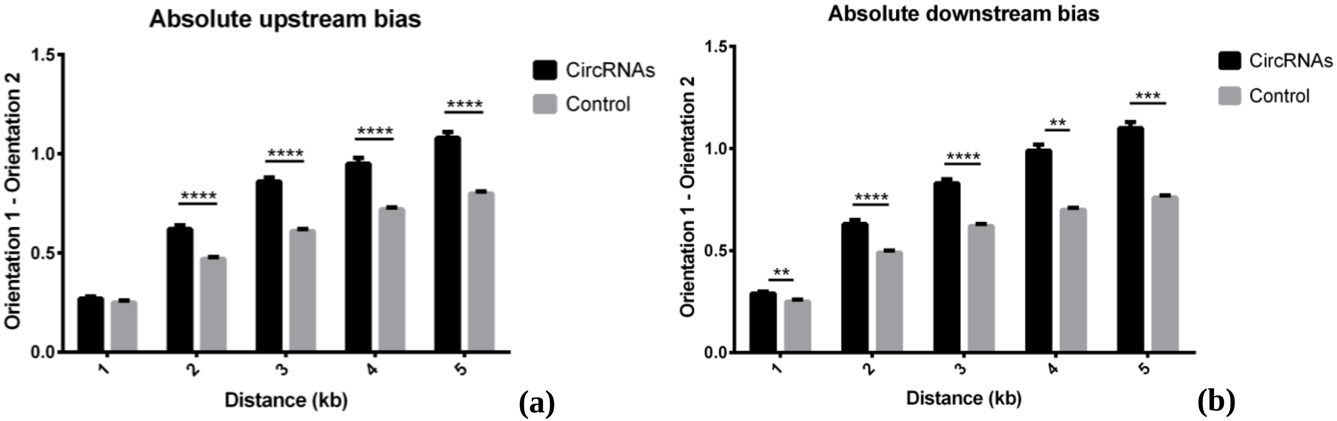


Figure 43. Average absolute orientation bias in upstream **(a)** and downstream **(b)** flanks. Plotted p-values: (****), $P < 10^{-4}$; (***), $P < 10^{-3}$; (**), $P < 10^{-2}$. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05).

We observed that circRNAs show a trend for Alu elements towards a preferred orientation in both upstream (**Figure 43a**) and downstream flanks (**Figure 43b**), which is significantly more accentuated than control.

Then, we calculated bias for each flank. A positive bias indicates a trend for Alu elements being in sense orientation, whereas a negative bias represents a trend for Alu elements being in antisense orientation.

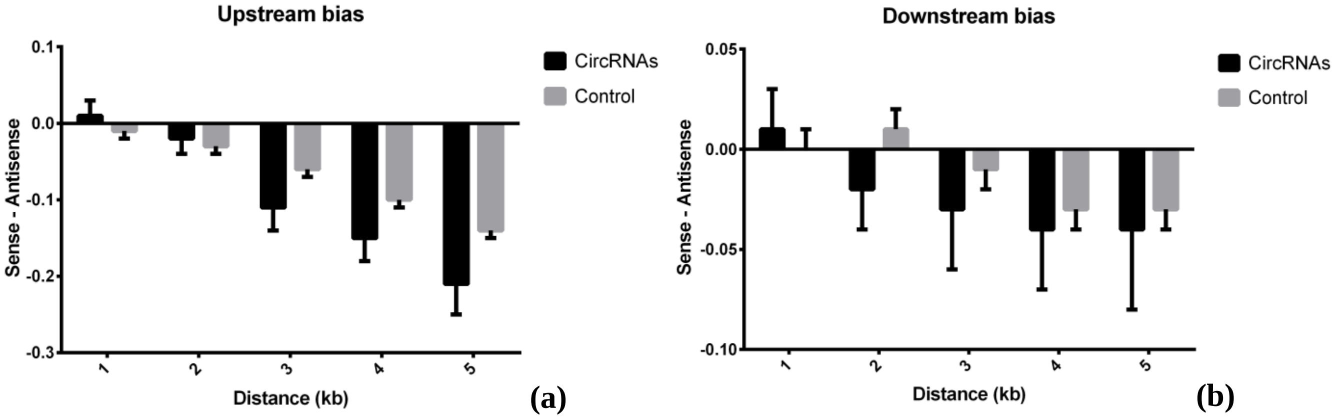


Figure 44. Average orientation bias in in upstream **(a)** and downstream **(b)** flanks. In both plots, circRNAs do not differ significantly from control for all distances (*n.s.*, $P > 0.05$).

Despite the absence of statistically significant differences between circRNAs and control, we observed that Alu elements in both datasets tend to be in antisense orientation (**Figure 44**), which may be explained by the overall abundance of antisense Alu elements flanking circRNAs from both datasets (**Figure 45**). However, we observed an increased bias in circRNAs for having antisense Alu elements, particularly in upstream regions (**Figure 44a**), while bias is much lower in downstream flanks (**Figure 44b**). Such an increased bias for antisense Alu elements in upstream regions of circRNAs may be explained by their genomic location in upstream exons near the 5'-UTR (**Appendix B, Table B1**), where it was demonstrated a preferential insertion of antisense Alu elements [163].

We tested the influence of the genomic location in the environment around circRNAs and evaluated editing in Alu elements flanking control with the same location as circRNAs (**Appendix B, Figures B1-B10**). The observed differences were towards the same conclusions, despite the observed their lower significance. Therefore, genomic location towards the 5'-UTR is not the only factor which may influence circRNA biogenesis.

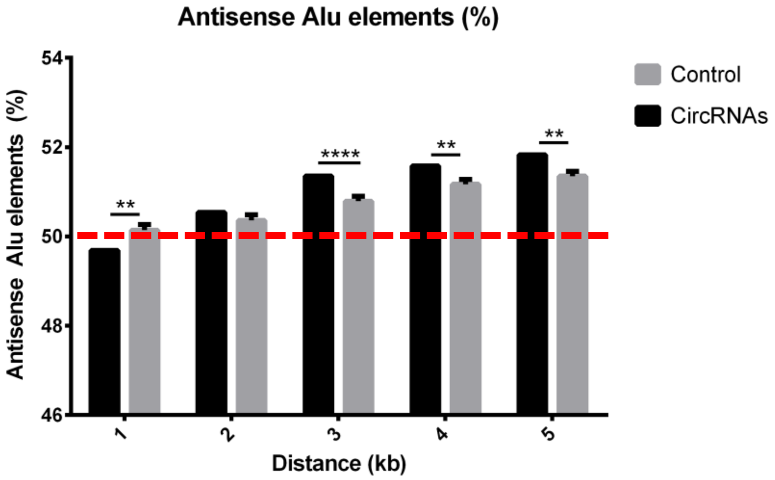


Figure 45. Abundance of Alu elements in antisense orientation. Plotted p-values: (****), $P < 10^{-4}$; (**), $P < 10^{-2}$. Unplotted p-values represent insignificant differences (*n.s.*, $P > 0.05$). Dashed horizontal line represents an overall abundance of 50%.

As mentioned previously, antisense Alu elements may constitute signals, such as polypyrimidine tracts derived from their poly(U) sequences, that activate new splice sites [106]. These Alu elements have been associated with regulation of alternative exons, especially when they are upstream to the regulated exon [101]. Considering the high splicing activity caused by Alu elements in the 5'-UTR of some genes which were identified in gene ontology analysis [105], these particular genes may create the ideal environment for alternative splicing with the presence of different splicing signaling sequences and the formation of secondary structures between Alu elements.

Bearing in mind that complementarity of the Alu pairs near the splice sites may significantly influence splicing efficiency by slowing down the action of the spliceosome [101], the reduced complementarity of inverted pairs near splice sites (**Figure 41**) may be a result of a higher regulation of the spliceosome, which could explain the increased specificity of ADARs (**Figure 40**).

Since the increased number of circRNAs with at least 1 edited Alu element (**Figure 39a**), as well as the insignificant differences in editing rate of Alu elements in close proximity to the splice sites (**Figure 39b**), ADARs could have a major influence in the stability of inverted pairs, promoting their formation or disruption. Therefore, A to I RNA editing may appear as a mechanism of regulating secondary structures that direct alternative splicing into circRNAs.

4. Conclusion and Future Directions

Several mechanisms regarding the regulation of splicing by secondary structures have been proposed. Two major mechanisms that were proposed for the occurrence of circularisation rely on long-range interactions between distant splice sites [164].

The first mechanism consists of direct backsplicing (**Figure 43a**), which is considered a form of splicing favored by intronic motifs which interact and bring both splice sites close together, resulting in a circular product without a linear counterpart [113]. The other possible mechanism consists of exon skipping, which is also promoted by RNA pairing between introns (**Figure 43b**). In the latter case, skipped exons would be looped out of the linear mRNA in order to approximate splice sites of the exons directly flanking these skipped exons [157, 158], resulting in a circRNA derived from skipped exons and an alternatively spliced linear mRNA.

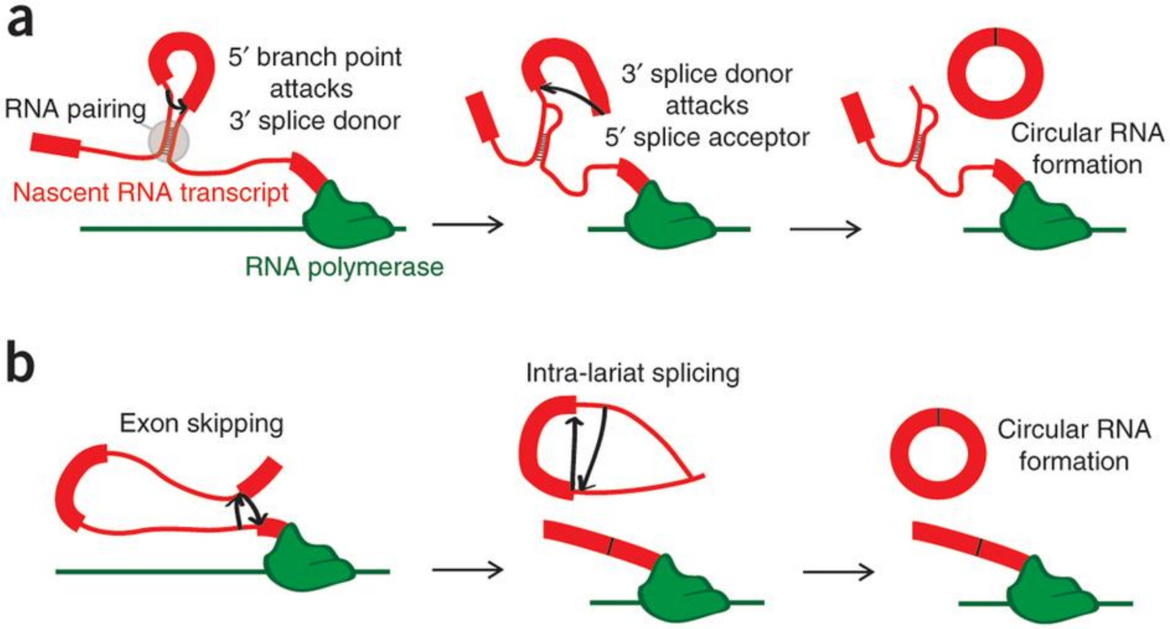


Figure 46. Proposed mechanisms of circRNA biogenesis [164].

Exon skipping appears as an appealing mechanism to explain the biogenesis of circRNAs. Miriami *et al.* analysed several genes reported to be affected by exon skipping and showed the enrichment of sequence motifs rich in G or C, which could base pair with complementary sequences and promote exon skipping or inclusion, depending if these pairs were constituted between introns or within the same, respectively (**Figure 47**) [165].

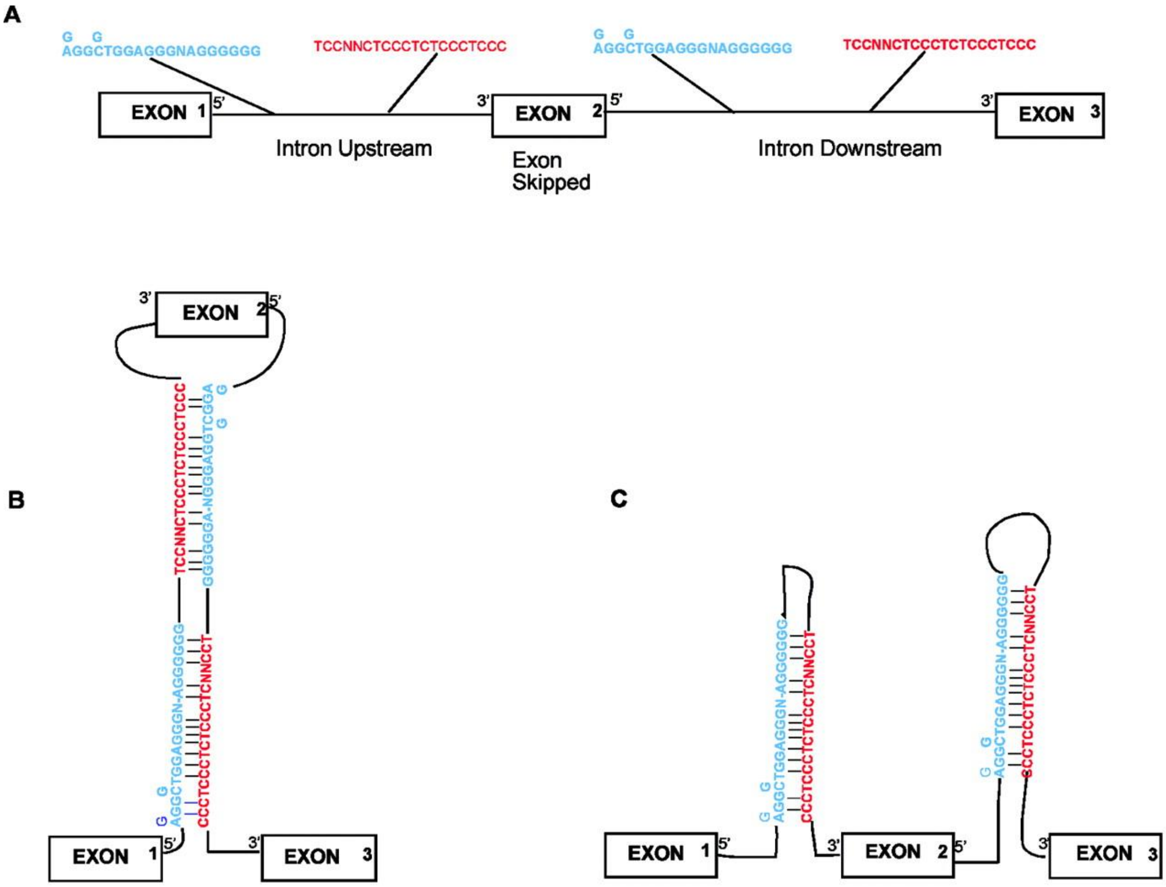


Figure 47. Influence of sequence motifs in regulation of alternative splicing. Base pairing between complementary motifs in different introns result in exon skipping **(b)**, while the same intron promote exon inclusion **(c)** [165].

These interactions generate different secondary structures that regulate alternative splicing, exposing or blocking splice sites of alternative exons. These C-rich and G-rich elements were shown to be conserved in other species [165], which mean that Alu elements are not the only sequences responsible for interactions between intronic regions that could promote alternative splicing.

Wong *et al.* showed the occurrence of alternative splicing of telomerase pre-mRNA transcripts caused by a variable number of short repeats flanking exon-intron junctions, which change the proximity of alternatively spliced exons and may expose target sites for spliceosomal components [166]. This mechanism could be applied to other eukaryotes, where circRNAs have been recently detected and characterised, some of them with short intronic flanking regions [151]. Inverted Alu pairs may promote increased circularisation in primates as these are highly stable structures which bring splice sites in close proximity.

The absence of miRNA and siRNA pathways in some eukaryotic organisms where circRNAs were identified [151] suggest another main function for circRNAs, rather than acting as competing endogenous RNAs [167]. A stronger hypothesis is that, due to their genomic location in upstream regions of genes, circRNAs could act as mRNA traps by sequestering the translation start site, leaving a noncoding transcript and therefore reduce protein concentration [164], which could explain the regulation of expression of some circRNAs under stress conditions [151]. In fact, circRNAs are indeed more prone to contain the initiation codon than our control (13.78% > 8.67% ± 0.04%, $P < 10^{-4}$), even when controlling the genomic location (13.78% > 9.37% ± 0.06%, $P < 10^{-4}$).

Therefore, circRNA biogenesis may be a result of highly regulated alternative splicing, resulting in stable circular structures that may control gene expression in several ways. We showed an enrichment of Alu elements flanking circRNAs, which may represent a major impact in their formation. We found that particular genes which were shown to have Alu elements with high splicing activity near the 5’-UTR were among those forming circRNAs. Moreover, an increased propensity for Alu elements being in antisense orientation constitutes another factor for alternative splicing in these exons, and may affect circularisation.

Although circRNAs have an increased probability to have inverted Alu pairs in their flanking regions, these Alu elements are less frequently edited. However, due to the high abundance of circRNAs with at least 1 edited Alu element and the increased specificity of ADARs in inverted Alu pairs near splice sites, we suggest that RNA editing may act as mechanism to regulate their stability. The created secondary structures may regulate the approximation of splice sites and binding of splicing factors that lead to circularisation. ADARs could intervene in gene expression by regulating the amount of generated circRNAs, thus its expression should lead to different expression of circRNAs and their respective genes.

For future plans, laboratory approaches are required to complement this computational approach and confirm our hypothesis. Thus, we need to determine if ADARs have a biological influence in circRNA biogenesis and then identify which proteins may be involved in circularisation *in vivo,* condition which provides the proteome complexity required for a better understanding of this process.

First, it is required to assess the involvement of inverted Alu pairs in circularisation. We could first create cell lines treated with siRNA to knockdown the expression of ADAR1 and ADAR2 and assess the expression of circRNAs through RT-PCR with outward primers after ribonuclease R digestion [113]. If the expression of circRNAs changed in the ADAR knockout cells compared to a control cell line expressing ADARs, we could confirm the influence of ADARs in circRNA biogenesis.

Then, to confirm the hypothesis of ADARs competing with splicing factors for the inverted Alu pairs, we could try two different but complementary approaches – the first by searching which proteins could bind to the inverted Alu pairs by tagging sequences near the closest pair, the second by searching for ligands of specific splicing factors.

Both of these approaches require ultraviolet radiation (UV) crosslinking to fix the contacts between naturally photoreactive RNA nucleosides and specific amino acids from RBPs [168].

We could build a construct expressing specific RNAs with inverted Alu pairs reported as circRNAs, with sequences flanking the closest pair that could be labelled to bind to streptavidin magnetic beads. These RNAs could be enriched for UV crosslinking by eluting only the tagged RNAs with bound RBPs, before washing to remove the streptavidin tags [169]. After ribonuclease treatment, these RBPs could be digested with trypsin and determined by mass spectrometry (MS) [168].

An alternative approach could be the application of crosslinking followed by immunoprecipitation (CLIP). This technique is based on the same principle of in vivo UV crosslinking in cells with ribonuclease digestion, however it is directed to find which RNAs are bound to a specific RBP by immunoprecipitating with antibodies which specifically recognise that protein [170]. These RNP complexes would be separated from free RBP through gel electrophoresis, followed by digestion with proteinase K to digest the RBP and allow reverse transcription of the RNA ligand, in order to obtain its sequence and map it in the genome [170]. We could test specific splicing factors such as PTB, hnRNP C, TIA-1 or HuR, which have been linked with U-rich sequences – which may be found in antisense Alu elements – and may affect splicing patterns [171].

Once identified, we could test the importance of these proteins in circularisation by knocking down their expression and assess the expression of circRNAs through RT-PCR with outward primers after ribonuclease R digestion [113]. If the influence of these proteins in circularisation is found to be important, we could additionally test the mRNA trap hypothesis as a potential function of circRNAs by inhibiting the formation of circRNAs and measure the respective protein expression levels.

An increased knowledge of these RNA structures about their biogenesis and on whether it only occurs in specific genes may provide us the tools to identify the potential biological functions of circRNAs and possibly to regulate RNA processing and gene expression, with potential biomedical applications.

REFERENCES

1. Naidoo, N., *et al.*, *Human genetics and genomics a decade after the release of the draft sequence of the human genome*. Human Genomics, 2011. **5**(6): pp. 577–622.
2. Lander, E. S., *et al.*, *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): pp. 860–921.
3. Venter, J. C., *et al.*, *The sequence of the human genome*. Science, 2001. **291**(5507): pp. 1304–1351.
4. Kim, M.-S., *et al.*, *A draft map of the human proteome*. Nature, 2014. **509**(7502): pp. 575–581.
5. Crick, F., *Central Dogma of Molecular Biology*. Nature, 1970. **227**(5258): pp. 561–563.
6. Graveley, B. R., *Alternative splicing: increasing diversity in the proteomic world*. Trends in Genetics, 2001. **17**(2): pp. 100–107.
7. Pullirsch, D., Jantsch, M. F., *Proteome diversification by adenosine to inosine RNA editing*. RNA Biology, 2010. **7**(2): pp. 205–212.
8. Gott, J. M., Emeson, R. B., *Functions and mechanisms of RNA editing*. Annual Review of Genetics, 2000. **34**: pp. 499–534.
9. Benne, R., *et al.*, *Major Transcript of the Frameshifted CoxII Gene from Trypanosome Mitochondria Contains 4 Nucleotides That Are Not Encoded in the DNA*. Cell, 1986. **46**(6): pp. 819–826.
10. Bass, B. L., *RNA editing by adenosine deaminases that act on RNA*. Annual Review of Biochemistry, 2002. **71**: pp. 817–846.
11. Conticello, S. G., *The AID/APOBEC family of nucleic acid mutators*. Genome Biology, 2008. **9**(6):
12. Wulff, B. E., Nishikura, K., *Substitutional A-to-I RNA editing*. Wiley Interdisciplinary Reviews RNA, 2010. **1**(1): pp. 90–101.
13. Hundley, H. A., Bass, B. L., *ADAR editing in double-stranded UTRs and other noncoding RNA sequences*. Trends in Biochemical Sciences, 2010. **35**(7): pp. 377–383.
14. Ling, J. Q., *et al.*, *Aminoacyl-tRNA Synthesis and Translational Quality Control*. Annual Review of Microbiology, 2009. **63**: pp. 61–78.

15. Phizicky, E. M., Hopper, A. K., *tRNA biology charges to the front*. Genes and Development, 2010. **24**(17): pp. 1832–1860.

16. Gustilo, E. M., *et al.*, *tRNA's modifications bring order to gene expression*. Current Opinion in Microbiology, 2008. **11**(2): pp. 134–140.

17. Motorin, Y., Helm, M., *tRNA Stabilization by Modified Nucleotides*. Biochemistry, 2010. **49**(24): pp. 4934–4944.

18. Nishikura, K., *Functions and Regulation of RNA Editing by ADAR Deaminases*. Annual Review of Biochemistry, 2010. **79**: pp. 321–349.

19. Barraud, P., Allain, F. H. T., *ADAR Proteins: Double-stranded RNA and Z-DNA Binding Domains*. Current Topics in Microbiology and Immunology, 2012. **353**: pp. 35–60.

20. Cho, D. S. C., *et al.*, *Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA*. Journal of Biological Chemistry, 2003. **278**(19): pp. 17093–17102.

21. Valente, L., Nishikura, K., *RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions*. Journal of Biological Chemistry, 2007. **282**(22): pp. 16054–16061.

22. Chen, C. X., *et al.*, *A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains*. RNA, 2000. **6**(5): pp. 755–767.

23. Enstero, M., *et al.*, *Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA*. Nucleic Acids Research, 2009. **37**(20): pp. 6916–6926.

24. Macbeth, M. R., *et al.*, *Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing*. Science, 2005. **309**(5740): pp. 1534–1539.

25. Ryter, J. M., Schultz, S. C., *Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA*. EMBO Journal, 1998. **17**(24): pp. 7505–7513.

26. Liu, Y., Samuel, C. E., *Mechanism of interferon action: Functionally distinct RNA-binding and catalytic domains in the interferon-inducible, double-stranded RNA-Specific adenosine deaminase*. Journal of Virology, 1996. **70**(3): pp. 1961–1968.

27. Barraud, P., *et al.*, *Solution structure of the N-terminal dsRBD of Drosophila ADAR and interaction studies with RNA*. Biochimie, 2012. **94**(7): pp. 1499–1509.

28. Fritz, J., *et al.*, *RNA-Regulated Interaction of Transportin-1 and Exportin-5 with the Double-Stranded RNA-Binding Domain Regulates Nucleocytoplasmic Shuttling of ADAR1*. Molecular and Cellular Biology, 2009. **29**(6): pp. 1487–1497.

29. Barraud, P., *et al.*, *A bimodular nuclear localization signal assembled via an extended double-stranded RNA-binding domain acts as an RNA-sensing signal for transportin 1*. Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(18): pp. E1852–E1861.

30. Athanasiadis, A., *et al.*, *The crystal structure of the Z beta domain of the RNA-editing enzyme ADAR1 reveals distinct conserved surfaces among Z-domains*. Journal of Molecular Biology, 2005. **351**(3): pp. 496–507.

31. Desterro, J. M. P., *et al.*, *Dynamic association of RNA-editing enzymes with the nucleolus*. Journal of Cell Science, 2003. **116**(9): pp. 1805–1818.

32. Keegan, L. P., *et al.*, *Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects*. Nucleic Acids Research, 2011. **39**(16): pp. 7249–7262.

33. Stefl, R., *et al.*, *The Solution Structure of the ADAR2 dsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove*. Cell, 2010. **143**(2): pp. 225–237.

34. Wahlstedt, H., Öhman, M., *Site-selective versus promiscuous A-to-I editing*. Wiley Interdisciplinary Reviews RNA, 2011. **2**(6): pp. 761–771.

35. Lehmann, K. A., Bass, B. L., *The importance of internal loops within RNA substrates of ADAR1*. Journal of Molecular Biology, 1999. **291**(1): pp. 1–13.

36. Tian, N., *et al.*, *A structural determinant required for RNA editing*. Nucleic Acids Research, 2011. **39**(13): pp. 5669–5681.

37. Rieder, L. E., *et al.*, *Tertiary structural elements determine the extent and specificity of messenger RNA editing*. Nature Communications, 2013. **4**: p. 2232.

38. Polson, A. G., Bass, B. L., *Preferential Selection of Adenosines for Modification by Double-Stranded-RNA Adenosine-Deaminase*. EMBO Journal, 1994. **13**(23): pp. 5701–5711.

39. Lehmann, K. A., Bass, B. L., *Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities*. Biochemistry, 2000. **39**(42): pp. 12875–12884.

40. Kuttan, A., Bass, B. L., *Mechanistic insights into editing-site specificity of ADARs*. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(48): pp. E3295–E3304.

41. Yi-Brunozzi, H. Y., *et al.*, *Conformational changes that occur during an RNA-editing adenosine deamination reaction*. Journal of Biological Chemistry, 2001. **276**(41): pp. 37827–37833.

42. Eggington, J. M., *et al.*, *Predicting sites of ADAR editing in double-stranded RNA*. Nature Communications, 2011. **2**(319):

43. Wilcox, J. L., Bevilacqua, P. C., *pK(a) Shifting in Double-Stranded RNA Is Highly Dependent upon Nearest Neighbors and Bulge Positioning*. Biochemistry, 2013. **52**(42): pp. 7470–7476.

44. Athanasiadis, A., *et al.*, *Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome*. PLoS Biology, 2004. **2**(12): p. e391.

45. Levanon, E. Y., *et al.*, *Systematic identification of abundant A-to-I editing sites in the human transcriptome*. Nature Biotechnology, 2004. **22**(8): pp. 1001–1005.

46. Li, M. Y., *et al.*, *Widespread RNA and DNA Sequence Differences in the Human Transcriptome*. Science, 2011. **333**(6038): pp. 53–58.

47. Ramaswami, G., *et al.*, *Accurate identification of human Alu and non-Alu RNA editing sites*. Nature Methods, 2012. **9**(6): pp. 579–581.

48. Peng, Z. Y., *et al.*, *Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome*. Nature Biotechnology, 2012. **30**(3): pp. 253–260.

49. Chen, L., *Characterization and comparison of human nuclear and cytosolic editomes*. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(29): pp. E2741–E2747.

50. Ramaswami, G., Li, J. B., *RADAR: a rigorously annotated database of A-to-I RNA editing*. Nucleic Acids Research, 2014. **42**(1): pp. D109–D113.

51. Pinto, Y., *et al.*, *Mammalian conserved ADAR targets comprise only a small fragment of the human editosome*. Genome Biology, 2014. **15**(1): p. R5.

52. Peixoto, A. A., Hall, J. C., *Analysis of temperature-sensitive mutants reveals new genes involved in the courtship song of Drosophila*. Genetics, 1998. **148**(2): pp. 827–838.

53. Hanrahan, C. J., *et al., RNA editing of the Drosophila para Na(+) channel transcript. Evolutionary conservation and developmental regulation*. Genetics, 2000. **155**(3): pp. 1149–1160.

54. Li, X., *et al., The ADAR RNA editing enzyme controls neuronal excitability in Drosophila melanogaster*. Nucleic Acids Research, 2013. **42**(2): pp. 1139–1151.

55. Bhogal, B., *et al., Modulation of dADAR-dependent RNA editing by the Drosophila fragile X mental retardation protein*. Nature Neuroscience, 2011. **14**(12): pp. 1517–1524.

56. Tonkin, L. A., *et al., RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans*. EMBO Journal, 2002. **21**(22): pp. 6025–6035.

57. Hotton III, N., *A Simplified Family Tree of Life*, in *The Evidence of Evolution*, Smithsonian, 1968.

58. Schoft, V. K., *et al., Regulation of glutamate receptor B pre-mRNA splicing by RNA editing*. Nucleic Acids Research, 2007. **35**(11): pp. 3723–3732.

59. Higuchi, M., *et al., Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2*. Nature, 2000. **406**(6791): pp. 78–81.

60. Ohlson, J., *et al., Editing modifies the GABA(A) receptor subunit alpha 3*. RNA, 2007. **13**(5): pp. 698–703.

61. Schmauss, C., Howe, J. R., *RNA Editing of Neurotransmitter Receptors in the Mammalian Brain*. Science STKE, 2002. **2002**(133): p. pe26.

62. Shtrichman, R., *et al., Altered A-to-I RNA Editing in Human Embryogenesis*. PLoS One, 2012. **7**(7): p. e41576.

63. XuFeng, R., *et al., ADAR1 is required for hematopoietic progenitor cell survival via RNA editing*. Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(42): pp. 17763–17768.

64. Osenberg, S., *et al., Alu Sequences in Undifferentiated Human Embryonic Stem Cells Display High Levels of A-to-I RNA Editing*. PLoS One, 2010. **5**(6): p. e11173.

65. Wahlstedt, H., *et al.*, *Large-scale mRNA sequencing determines global regulation of RNA editing during brain development*. Genome Research, 2009. **19**(6): pp. 978–986.

66. Laurencikiene, J., *et al.*, *RNA editing and alternative splicing: the importance of co-transcriptional coordination*. EMBO Reports, 2006. **7**(3): pp. 303–307.

67. Raitskin, O., *et al.*, *RNA editing activity is associated with splicing factors in lnRNP particles: The nuclear pre-mRNA processing machinery.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(12): pp. 6571–6576.

68. Riedmann, E. M., Jantsch, M. F., *An editor controlled by transcription*. EMBO Reports, 2006. **7**(3): pp. 269–270.

69. Nevo-Caspi, Y., *et al.*, *A-to-I RNA Editing Is Induced Upon Hypoxia*. Shock, 2011. **35**(6): pp. 585–589.

70. Zhang, Z., Carmichael, G. G., *The fate of dsRNA in the Nucleus: A p54nrb-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs*. Cell, 2001. **106**(4): pp. 465–475.

71. Bond, C. S., Fox, A. H., *Paraspeckles: nuclear bodies built on long noncoding RNA.* Journal of Cell Biology, 2009. **186**(5): pp. 637–644.

72. Ben-Zvi, M., *et al.*, *F11R Expression upon Hypoxia Is Regulated by RNA Editing*. PLoS One, 2013. **8**(10): p. e77702.

73. Zhou, J., *et al.*, *On the mechanism of induction of heterochromatin by the RNA-binding protein vigilin.* RNA, 2008. **14**(9): pp. 1773–1781.

74. Wang, Q., *et al.*, *Vigilins bind to promiscuously A-to-I-edited RNAs and are involved in the formation of heterochromatin*. Current Biology, 2005. **15**(4): pp. 384–391.

75. Nie, Y. Z., *et al.*, *ADAR1 interacts with NF90 through double-stranded RNA and regulates NF90-mediated gene expression independently of RNA editing*. Molecular and Cellular Biology, 2005. **25**(16): pp. 6956–6963.

76. DeCerbo, J., Carmichael, G. G., *Retention and repression: Fates of hyperedited RNAs in the nucleus*. Current Opinion in Cell Biology, 2005. **17**(3): pp. 302–308.

77. Scadden, A. D. J., *The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage*. Nature Structural and Molecular Biology, 2005. **12**(6): pp. 489–496.

78. Sarvestani, S. T., *et al.*, *Inosine-Mediated Modulation of RNA Sensing by Toll-Like Receptor 7 (TLR7) and TLR8.* Journal of Virology, 2014. **88**(2): pp. 799–810.

79. George, C. X., *et al.*, *Adenosine deaminases acting on RNA, RNA editing, and interferon action*. Journal of Interferon and Cytokine Research, 2011. **31**(1): pp. 99–117.

80. Samuel, C. E., *Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral*. Virology, 2011. **411**(2): pp. 180–193.

81. Chen, R. X., *et al.*, *RNA Editing and its Control in Hepatitis Delta Virus Replication*. Viruses, 2010. **2**(1): pp. 131–146.

82. Nishikura, K., *Editor meets silencer: crosstalk between RNA editing and RNA interference*. Nature Reviews Molecular Cell Biology, 2006. **7**(12): pp. 919–931.

83. Ota, H., *et al.*, *ADAR1 Forms a Complex with Dicer to Promote MicroRNA Processing and RNA-Induced Gene Silencing*. Cell, 2013. **153**(3): pp. 575–589.

84. Hartner, J. C., *et al.*, *ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling.* Nature Immunology, 2009. **10**(1): pp. 109–115.

85. Baldwin, J. M., *A New Factor in Evolution*. The American Naturalist, 1896. **30**(354): pp. 441–451, 536–553.

86. Simpson, G. G., *The Baldwin effect*. Evolution, 1953. **7**: pp. 110–117.

87. Ancel, L. W., *A quantitative model of the Simpson-Baldwin Effect.* Journal of Theoretical Biology, 1999. **196**(2): pp. 197–209.

88. Gommans, W. M., *et al.*, *RNA editing: a driving force for adaptive evolution?* Bioessays, 2009. **31**(10): pp. 1137–1145.

89. Jin, Y., *et al.*, *Origins and evolution of ADAR-mediated RNA editing.* IUBMB Life, 2009. **61**(6): pp. 572–578.

90. Batzer, M. A., Deininger, P. L., *Alu repeats and human genomic diversity.* Nature Reviews Genetics, 2002. **3**(5): pp. 370–379.

91. Ichiyanagi, K., *Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs*. Genes and Genetic Systems, 2013. **88**(1): pp. 19–29.

92. Walters, R. D., *et al.*, *InvAluable Junk: The Cellular Impact and Function of Alu and B2 RNAs*. IUBMB Life, 2009. **61**(8): pp. 831–837.

93. Hasler, J., *et al.*, *Useful 'junk': Alu RNAs in the human transcriptome*. Cellular and Molecular Life Sciences, 2007. **64**(14): pp. 1793–1800.

94. Pandey, R., *et al.*, *Heat shock factor binding in Alu repeats expands its involvement in stress through an antisense mechanism*. Genome Biology, 2011. **12**(11): p. R117.

95. Grover, D., *et al.*, *Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition*. Bioinformatics, 2004. **20**(6): pp. 813–817.

96. Hasler, J., Strub, K., *Alu elements as regulators of gene expression*. Nucleic Acids Research, 2006. **34**(19): pp. 5491–5497.

97. Chen, C. J., *et al.*, *Using Alu Elements as Polyadenylation Sites: A Case of Retroposon Exaptation*. Molecular Biology and Evolution, 2009. **26**(2): pp. 327–334.

98. Smalheiser, N. R., Torvik, V. I., *Alu elements within human mRNAs are probable microRNA targets*. Trends in Genetics, 2006. **22**(10): pp. 532–536.

99. Fitzpatrick, T., Huang, S., *3'-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation*. Nucleus, 2012. **3**(4): pp. 359–369.

100. Gong, C. G., Maquat, L. E., *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements*. Nature, 2011. **470**(7333): pp. 284–288.

101. Lev-Maor, G., *et al.*, *Intronic Alus influence alternative splicing*. PLoS Genetics, 2008. **4**(9): p. e1000204.

102. Gal-Mark, N., *et al.*, *Alternative splicing of Alu exons - two arms are better than one*. Nucleic Acids Research, 2008. **36**(6): pp. 2012–2023.

103. Schmitz, J., Brosius, J., *Exonization of transposed elements: A challenge and opportunity for evolution*. Biochimie, 2011. **93**(11): pp. 1928–1934.

104. Lev-Maor, G., *et al.*, *RNA-editing-mediated exon evolution*. Genome Biology, 2007. **8**(2): p. R29.

105. Shen, S., *et al.*, *Widespread establishment and regulatory impact of Alu exons in human genes.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(7): pp. 2837–2842.

106. Zarnack, K., *et al.*, *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements*. Cell, 2013. **152**(3): pp. 453–466.

107. Rino, J., Carmo-Fonseca, M., *The spliceosome: a self-organized macromolecular machine in the nucleus?* Trends in Cell Biology, 2009. **19**(8): pp. 375–384.

108. Daniel, C., *et al.*, *Alu elements shape the primate transcriptome by cis-regulation of RNA editing*. Genome Biology, 2014. **15**: p. R28.

109. Salzman, J., *et al.*, *Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types*. PLoS One, 2012. **7**(2): p. e30733.

110. Xu, A. N., *et al.*, *Identification of a novel circularized transcript of the AML1 gene*. BMB Reports, 2013. **46**(3): pp. 163–168.

111. Danan, M., *et al.*, *Transcriptome-wide discovery of circular RNAs in Archaea*. Nucleic Acids Research, 2012. **40**(7): pp. 3131–3142.

112. Memczak, S., *et al.*, *Circular RNAs are a large class of animal RNAs with regulatory potency*. Nature, 2013. **495**(7441): pp. 333–338.

113. Jeck, W. R., *et al.*, *Circular RNAs are abundant, conserved, and associated with ALU repeats*. RNA, 2013. **19**(2): pp. 141–157.

114. Hentze, M. W., Preiss, T., *Circular RNAs: splicing's enigma variations*. EMBO Journal, 2013. **32**(7): pp. 923–925.

115. Hansen, T. B., *et al.*, *Natural RNA circles function as efficient microRNA sponges*. Nature, 2013. **495**(7441): pp. 384–388.

116. Hansen, T. B., *et al.*, *miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA*. EMBO Journal, 2011. **30**(21): pp. 4414–4422.

117. Chen, C. Y., Sarnow, P., *Initiation of Protein-Synthesis by the Eukaryotic Translational Apparatus on Circular RNAs*. Science, 1995. **268**(5209): pp. 415–417.

118. Hellen, C. U., Sarnow, P., *Internal ribosome entry sites in eukaryotic mRNA molecules*. Genes & development, 2001. **15**(13): pp. 1593–1612.

119. Van Rossum, G., *Python 2.7.6 Documentation*, *Python Software Foundation*, 1991. [Online]. Available: https://docs.python.org/2. [Accessed: April 23rd, 2014].

120. Oliphant, T. E., *Python for Scientific Computing*. Computing in Science and Engineering, 2007. **9**: pp. 10–20.

121. Moreira, W., *et al.*, *RPy2: A simple and efficient access to R from Python*, 2004. [Online]. Available: http://rpy.sourceforge.net/rpy2.html. [Accessed: April 23rd, 2014].

122. R Development Core Team, *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, 2008. [Online]. Available: http://www.r-project.org/. [Accessed: April 22nd, 2014].

123. Cock, P. J. A., *et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics (Oxford, England), 2009. **25**(11): pp. 1422–1423.

124. Smith, T. F., Waterman, M. S., *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **147**(1): pp. 195–197.

125. Rice, P., *et al., EMBOSS: the European Molecular Biology Open Software Suite.* Trends in Genetics, 2000. **16**(6): pp. 276–277.

126. Kent, W. J., *et al., The Human Genome Browser at UCSC*. Genome Research, 2002. **12**(6): pp. 996–1006.

127. Kruskal, W. H., Wallis, W. A., *Use of Ranks in One-Criterion Variance Analysis*. Journal of the American Statistical Association, 1952. **47**(260): pp. 583–621.

128. Mann, H., Whitney, D., *On a test of wether one of two random variables is stochastically larger than the other*. Annals of Mathematical Statistics, 1947. **18**(1): pp. 50–60.

129. Bonferroni, C. E., *Il calcolo delle assicurazioni su gruppi di teste*, in *Studi in Onore del Professore Salvatore Ortu Carboni*, Rome, 1935, pp. 13–60.

130. Huang, D. W., *et al., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nature Protocols, 2009. **4**(1): pp. 44–57.

131. Smit, A. F. A., *et al., RepeatMasker Open-4.0.3, Institute for Systems Biology*, 1996. [Online]. Available: http://www.repeatmasker.org. [Accessed: April 23rd, 2014].

132. Cohen, J., *A power primer*. Psychological Bulletin, 1992. **112**(1): pp. 155–159.

133. Carmi, S., *et al., Identification of widespread ultra-edited human RNAs*. PLoS Genetics, 2011. **7**(10): p. e1002317.

134. Student, *Probable Error of a Correlation Coefficient*. Biometrika, 1908. **6**(1): pp. 1–25.

135. Shapiro, S., Wilk, M., *An approximate analysis of variance test for normality*. Biometrika, 1965. **52**(3)–(4): pp. 591–611.

136. Razali, N. M., Wah, Y. B., *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests*. Journal of Statistical Modeling and Analytics, 2011. **2**(1): pp. 21–33.

137. Zhang, Y., *et al.*, *Circular Intronic Long Noncoding RNAs*. Molecular Cell, 2013. **51**(6): pp. 792–806.

138. Ozsolak, F., Milos, P. M., *RNA sequencing: advances, challenges and opportunities*. Nature Reviews Genetics, 2011. **12**(2): pp. 87–98.

139. Wang, K., *et al.*, *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery*. Nucleic Acids Research, 2010. **38**(18): p. e178.

140. Bolisetty, M. T., Beemon, K. L., *Splicing of internal large exons is defined by novel cis-acting sequence elements*. Nucleic Acids Research, 2012. **40**(18): pp. 9244–9254.

141. Shepard, P. J., Hertel, K. J., *Conserved RNA secondary structures promote alternative splicing*. RNA, 2008. **14**(8): pp. 1463–1469.

142. Hunter, S., *et al.*, *InterPro in 2011: New developments in the family and domain prediction database*. Nucleic Acids Research, 2012. **40**(D1): pp. D306–D312.

143. Takeda, J. I., *et al.*, *Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56 419 completely sequenced and manually annotated full-length cDNAs*. Nucleic Acids Research, 2006. **34**(14): pp. 3917–3928.

144. Ogata, H., *et al.*, *KEGG: Kyoto encyclopedia of genes and genomes*. Nucleic Acids Research, 1999. **27**(1): pp. 29–34.

145. Amm, I., *et al.*, *Protein quality control and elimination of protein waste: The role of the ubiquitin-proteasome system*. Biochimica et Biophysica Acta, 2014. **1843**(1): pp. 182–196.

146. Kleiger, G., Mayor, T., *Perilous journey: a tour of the ubiquitin-proteasome system*. Trends in Cell Biology, 2014. **24**(6): pp. 352–359.

147. Callahan, K. P., Butler, J. S., *TRAMP complex enhances RNA degradation by the nuclear exosome component Rrp6*. Journal of Biological Chemistry, 2010. **285**(6): pp. 3540–3547.

148. Collart, M. A., Panasenko, O. O., *The Ccr4-Not complex*. Gene, 2012. **492**(1): pp. 42–53.

149. Nagarajan, V. K., *et al.*, *XRN 5'→3' exoribonucleases: structure, mechanisms and functions*. Biochimica et Biophysica Acta, 2013. **1829**(6)–(7): pp. 590–603.

150. Ostareck, D. H., *et al.*, *DDX6 and its orthologs as modulators of cellular and viral RNA expression*. Wiley Interdisciplinary Reviews RNA, 2014. **Not issued**:

151. Wang, P. L., *et al.*, *Circular RNA Is Expressed across the Eukaryotic Tree of Life*. PLoS One, 2014. **9**(3): p. e90859.

152. Garncarz, W., *et al.*, *A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing*. RNA Biology, 2013. **10**(2): pp. 192–204.

153. Tariq, A., *et al.*, *RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation*. Nucleic Acids Research, 2013. **41**(4): pp. 2581–2593.

154. Solomon, O., *et al.*, *Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR)*. RNA, 2013. **19**(5): pp. 591–604.

155. Wan, Y., *et al.*, *Understanding the transcriptome through RNA structure*. Nature Reviews Genetics, 2011. **12**(9): pp. 641–655.

156. Wan, Y., *et al.*, *Landscape and variation of RNA secondary structure across the human transcriptome.* Nature, 2014. **505**(7485): pp. 706–709.

157. Warf, M. B., Berglund, J. A., *Role of RNA structure in regulating pre-mRNA splicing*. Trends in Biochemical Sciences, 2009. **35**(3): pp. 169–178.

158. Jin, Y., *et al.*, *New insights into RNA secondary structure in the alternative splicing of pre-mRNAs*. RNA Biology, 2011. **8**(3): pp. 450–457.

159. Babendure, J. R., *et al.*, *Control of mammalian translation by mRNA structure near caps*. RNA, 2006. **12**(5): pp. 851–861.

160. Shukla, S., Oberdoerffer, S., *Co-transcriptional regulation of alternative pre-mRNA splicing*. Biochimica et Biophysica Acta, 2012. **1819**(7): pp. 673–683.

161. Veloso, A., *et al.*, *Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications*. Genome Research, 2014. **24**(6): pp. 896–905.

162. Dujardin, G., *et al.*, *How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping*. Molecular Cell, 2014. **54**(4): pp. 683–690.

163. Linker, S., Hedges, D., *Linear decay of retrotransposon antisense bias across genes is contingent upon tissue specificity*. PLoS One, 2013. **8**(11): p. e79402.

164. Jeck, W. R., Sharpless, N. E., *Detecting and characterizing circular RNAs*. Nature Biotechnology, 2014. **32**(5): pp. 453–461.

165. Miriami, E., *et al.*, *Conserved sequence elements associated with exon skipping*. Nucleic Acids Research, 2003. **31**(7): pp. 1974–1983.

166. Wong, M. S., *et al.*, *Regulation of human telomerase splicing by RNA:RNA pairing*. Nature Communications, 2014. **5**: p. 3306.

167. Tay, Y., *et al.*, *The multilayered complexity of ceRNA crosstalk and competition.* Nature, 2014. **505**(7483): pp. 344–352.

168. Castello, A., *et al.*, *System-wide identification of RNA-binding proteins by interactome capture*. Nature Protocols, 2013. **8**(3): pp. 491–500.

169. Slobodin, B., Gerst, J. E., *A novel mRNA affinity purification technique for the identification of interacting proteins and transcripts in ribonucleoprotein complexes*. RNA, 2010. **16**(11): pp. 2277–2290.

170. Huppertz, I., *et al.*, *iCLIP: Protein-RNA interactions at nucleotide resolution*. Methods, 2014. **65**(3): pp. 274–287.

171. Chen, M., Manley, J. L., *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches*. Nature Reviews Molecular Cell Biology, 2009. **10**(11): pp. 741–754.

APPENDICES

A. *Flowcharts*

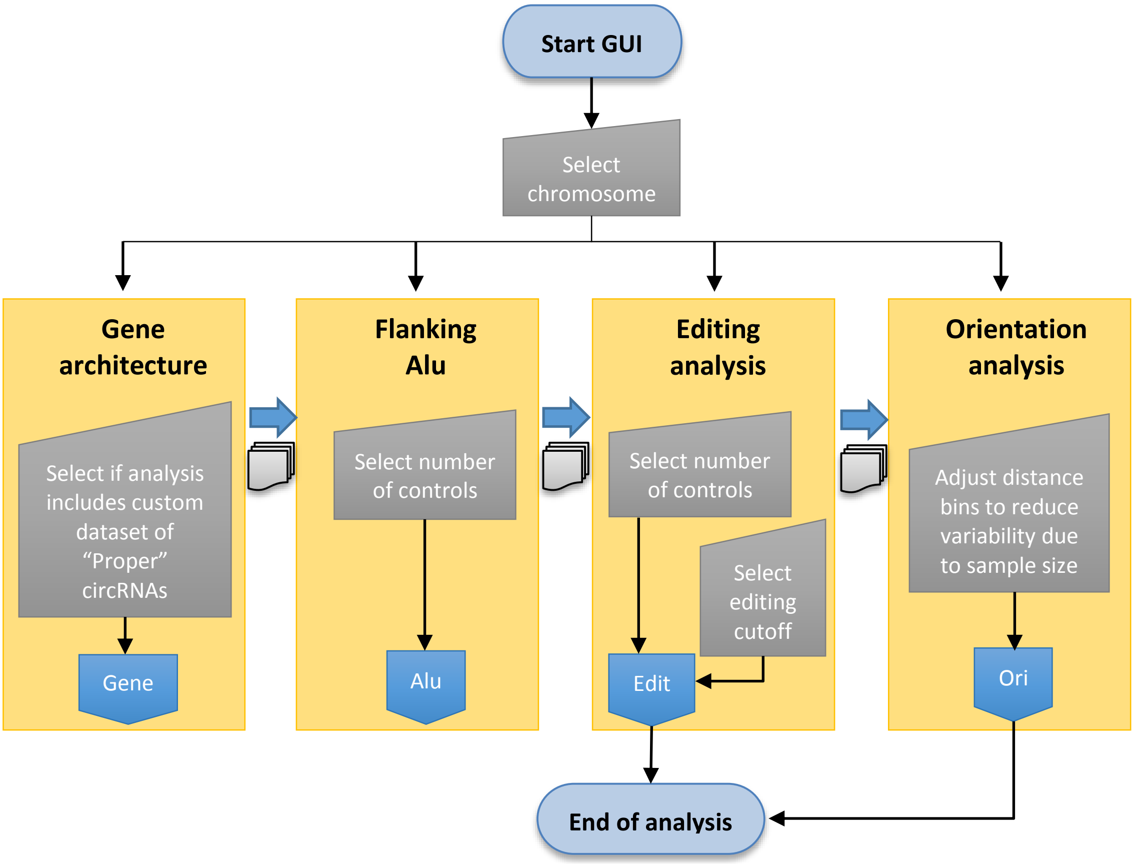
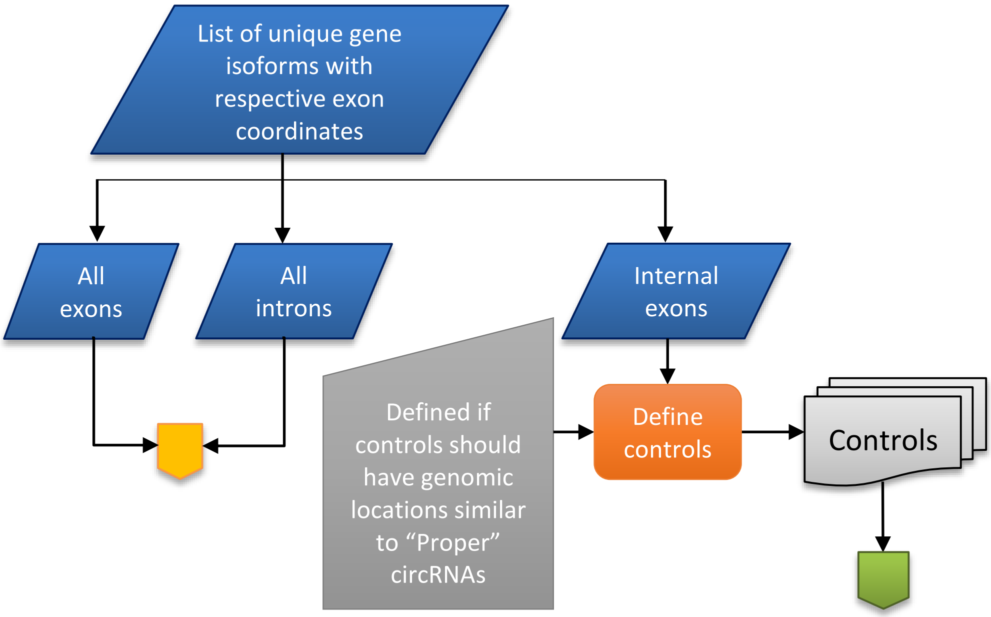

Start GUI
Select chromosome
Gene architecture
Select if analysis includes custom dataset of "Proper" circRNAs
Gene
Flanking Alu
Select number of controls
Alu
Editing analysis
Select number of controls
Select editing cutoff
Edit
Orientation analysis
Adjust distance bins to reduce variability due to sample size
Ori
End of analysis

Figure A1. Representation of the graphical user interface (GUI) which manages the multiple analyses that were performed in this work.


List of unique gene isoforms with respective exon coordinates
All exons
All introns
Internal exons
Defined if controls should have genomic locations similar to "Proper" circRNAs
Define controls
Controls

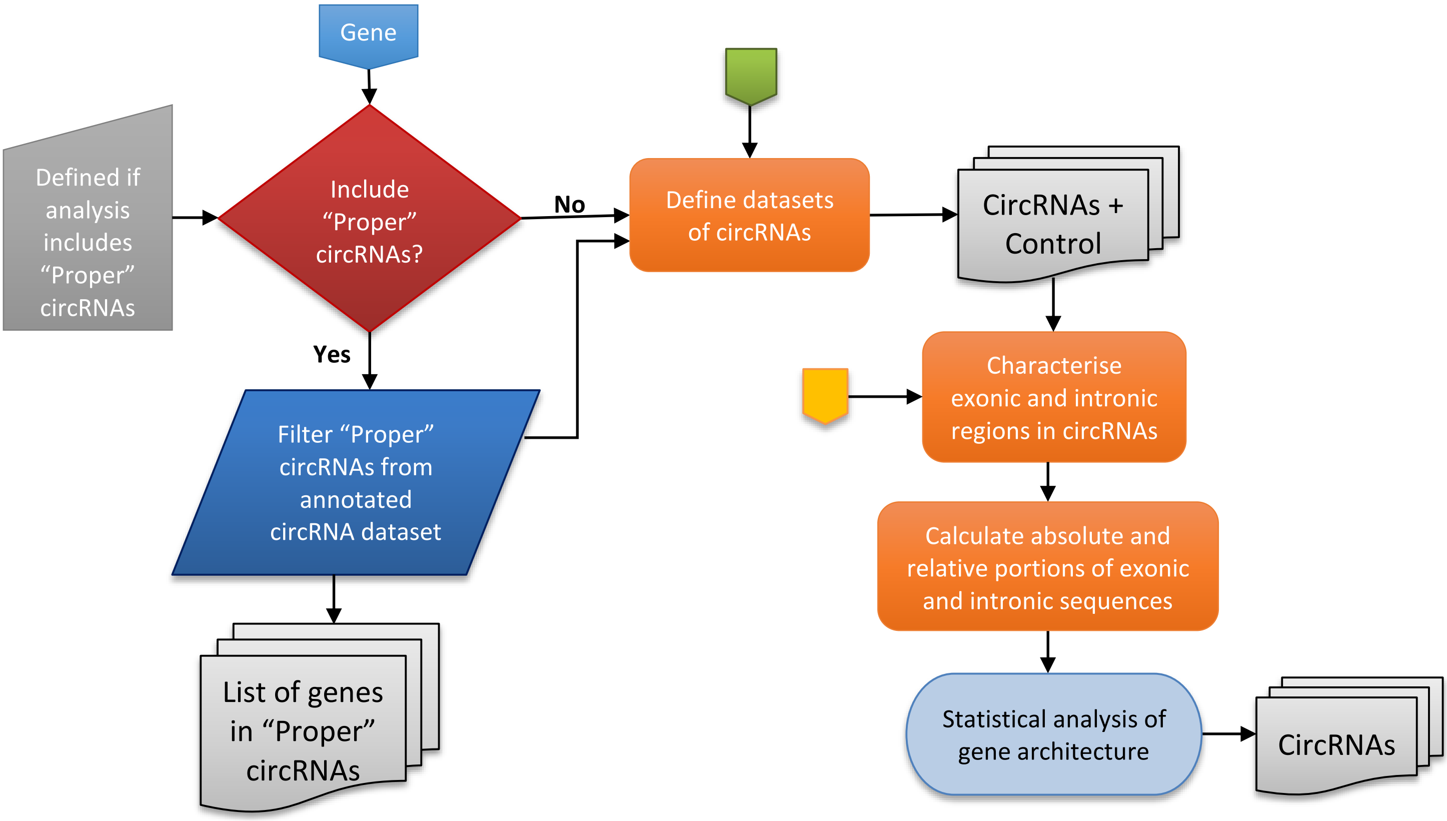
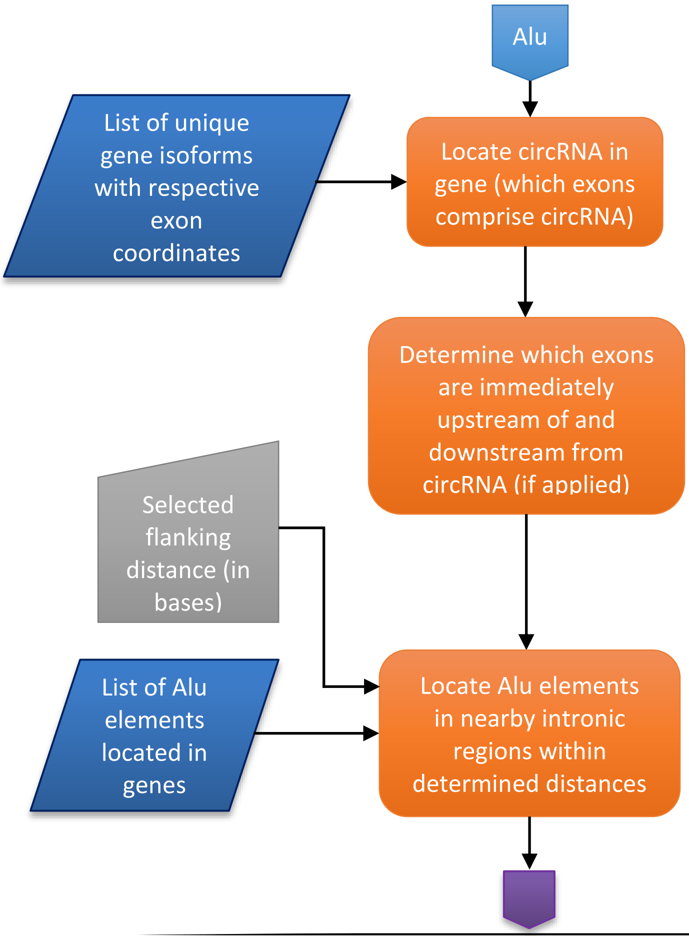

Gene
Defined if analysis includes "Proper" circRNAs
Include "Proper" circRNAs?
No
Yes
Define datasets of circRNAs
CircRNAs + Control
Filter "Proper" circRNAs from annotated circRNA dataset
List of genes in "Proper" circRNAs
Characterise exonic and intronic regions in circRNAs
Calculate absolute and relative portions of exonic and intronic sequences
Statistical analysis of gene architecture
CircRNAs

Figure A2. Flowchart representation of the *Gene architecture* process, derived from Figure A1.


Alu
List of unique gene isoforms with respective exon coordinates
Locate circRNA in gene (which exons comprise circRNA)
Determine which exons are immediately upstream of and downstream from circRNA (if applied)
Selected flanking distance (in bases)
List of Alu elements located in genes
Locate Alu elements in nearby intronic regions within determined distances

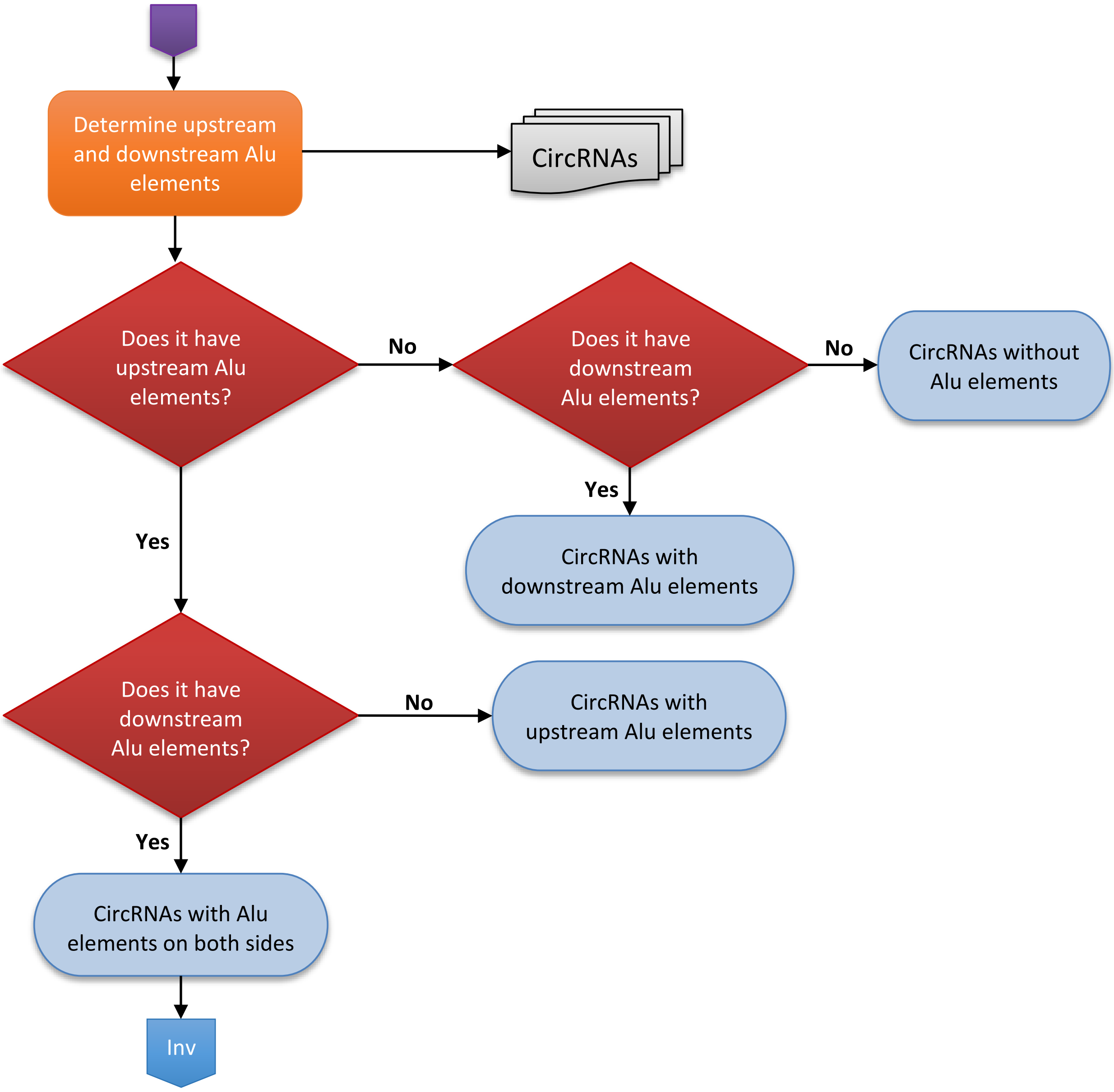

Determine upstream and downstream Alu elements
CircRNAs
Does it have upstream Alu elements?
No
Does it have downstream Alu elements?
No
CircRNAs without Alu elements
Yes
CircRNAs with downstream Alu elements
Yes
Does it have downstream Alu elements?
No
CircRNAs with upstream Alu elements
Yes
CircRNAs with Alu elements on both sides
Inv

Figure A3. Flowchart representation of the first part of the *Flanking Alu* process, derived from Figure A1.

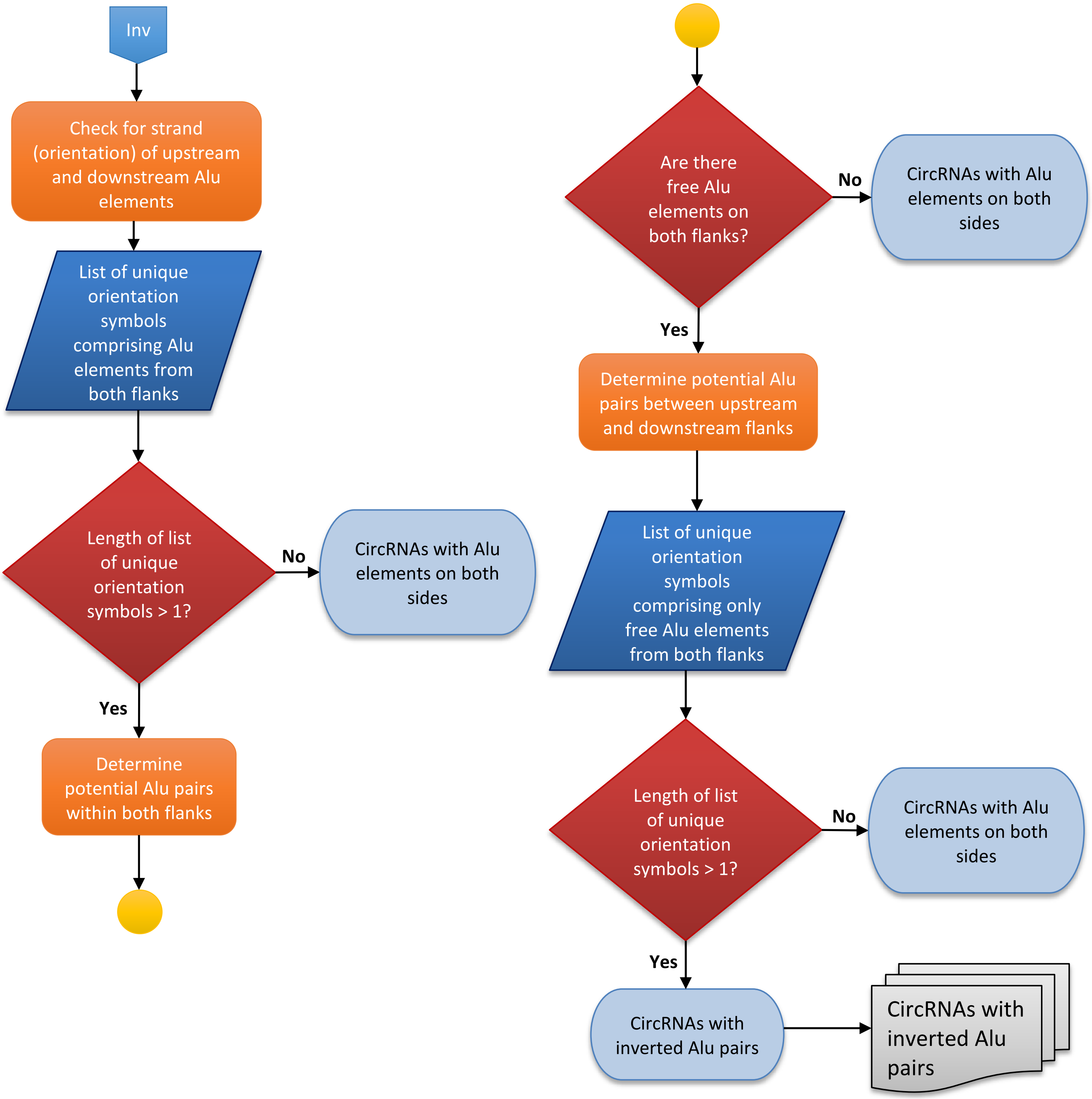

Inv
Check for strand (orientation) of upstream and downstream Alu elements
List of unique orientation symbols comprising Alu elements from both flanks
Length of list of unique orientation symbols > 1?
No
CircRNAs with Alu elements on both sides
Yes
Determine potential Alu pairs within both flanks
Are there free Alu elements on both flanks?
No
CircRNAs with Alu elements on both sides
Yes
Determine potential Alu pairs between upstream and downstream flanks
List of unique orientation symbols comprising only free Alu elements from both flanks
Length of list of unique orientation symbols > 1?
No
CircRNAs with Alu elements on both sides
Yes
CircRNAs with inverted Alu pairs
CircRNAs with inverted Alu pairs

Figure A4. Flowchart representation of the second part of the *Flanking Alu* process (extension of Figure A3), derived from Figure A1.

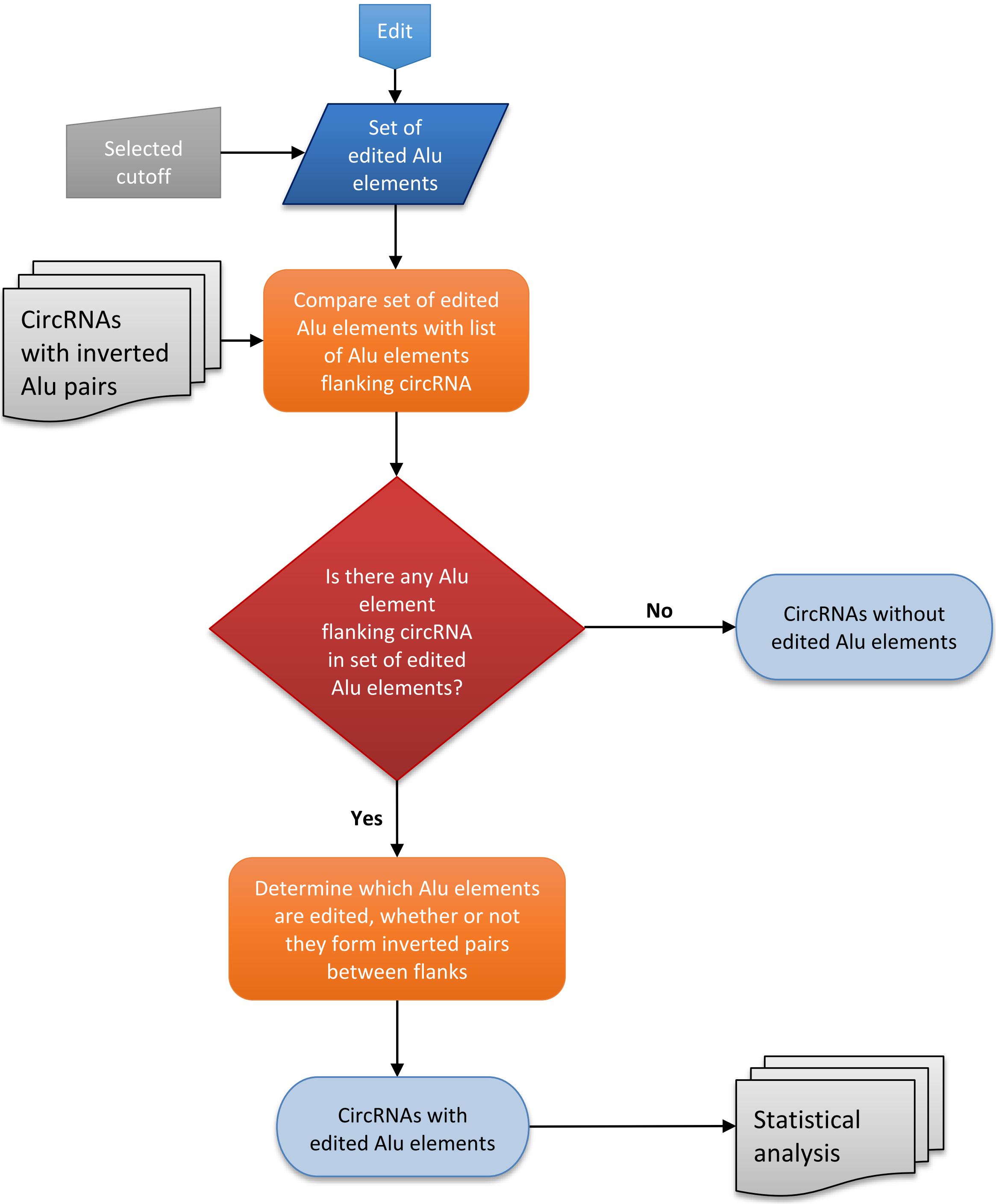

Edit
Selected cutoff
Set of edited Alu elements
CircRNAs with inverted Alu pairs
Compare set of edited Alu elements with list of Alu elements flanking circRNA
Is there any Alu element flanking circRNA in set of edited Alu elements?
No
CircRNAs without edited Alu elements
Yes
Determine which Alu elements are edited, whether or not they form inverted pairs between flanks
CircRNAs with edited Alu elements
Statistical analysis

Figure A5. Flowchart representation of the *Editing analysis* process, derived from Figure A1.

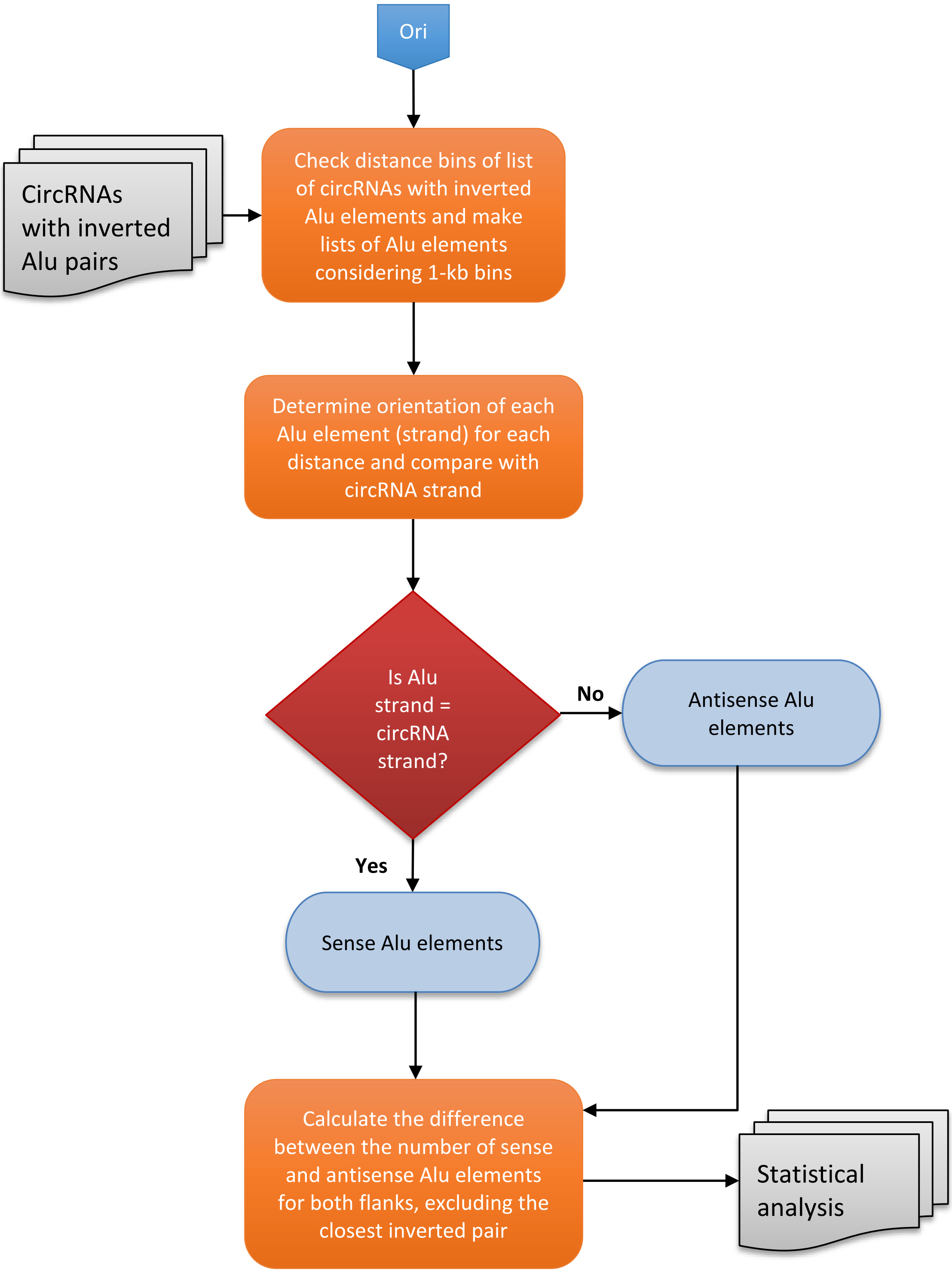

Ori
CircRNAs with inverted Alu pairs
Check distance bins of list of circRNAs with inverted Alu elements and make lists of Alu elements considering 1-kb bins
Determine orientation of each Alu element (strand) for each distance and compare with circRNA strand
Is Alu strand = circRNA strand?
No
Antisense Alu elements
Yes
Sense Alu elements
Calculate the difference between the number of sense and antisense Alu elements for both flanks, excluding the closest inverted pair
Statistical analysis

Figure A6. Flowchart representation of the *Orientation analysis* process, derived from Figure A1.

B. *Comparison between circRNAs and control sequences with the same location*

Table B1. Location of circRNAs relatively to their respective genes (exon numbers).

Exon number	Number of circRNAs that start at exon	Number of circRNAs that end at exon	Exon number	Number of circRNAs that start at exon	Number of circRNAs that end at exon
1	0	0	**14**	95	174
2	1494	142	**15**	77	122
3	953	464	**16**	93	135
4	643	648	**17**	68	98
5	492	666	**18**	60	84
6	388	642	**19**	34	66
7	300	539	**20**	41	48
8	288	471	**21**	40	72
9	230	397	**22**	40	50
10	171	303	**23**	30	47
11	142	254	**24**	24	37
12	139	236	**25**	28	30
13	126	195	**> 25**	188	264

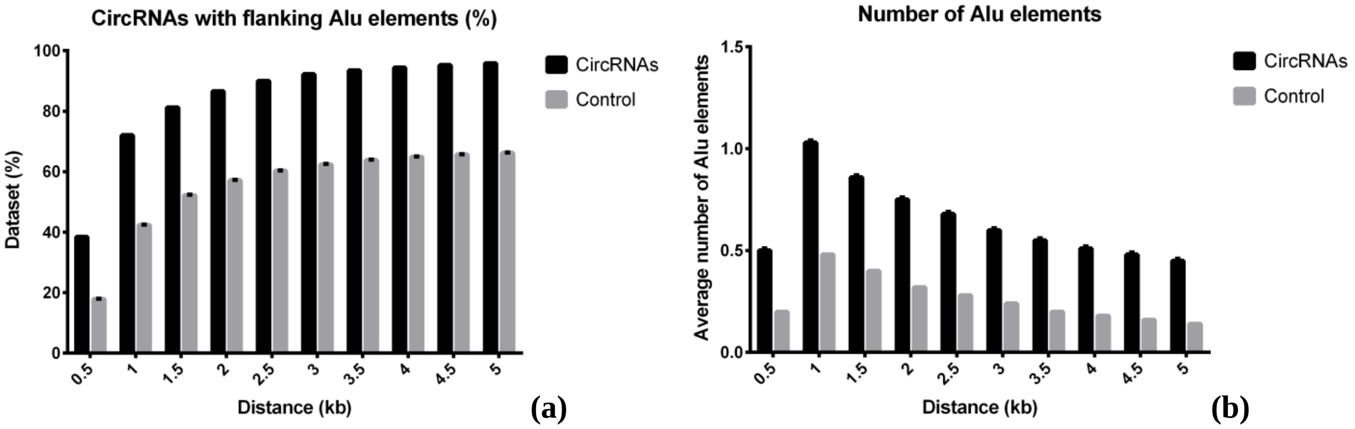


Figure B1. Cumulative number of circRNAs which are flanked by Alu elements **(a)** and average number of Alu elements flanking circRNAs **(b)** by flanking distance. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

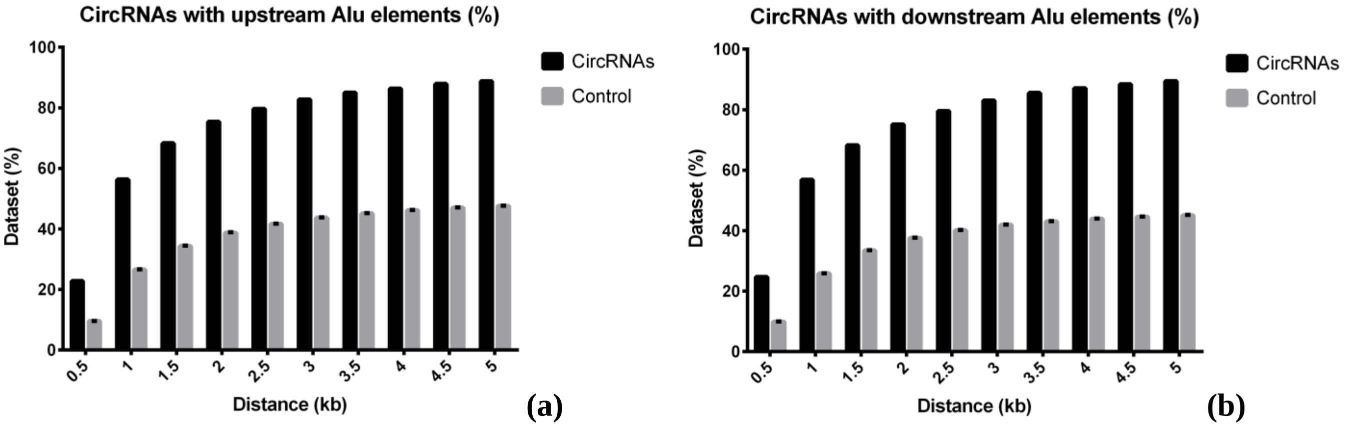


Figure B2. Cumulative number of circRNAs which are flanked by upstream **(a)** and downstream Alu elements **(b)** by flanking distance. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

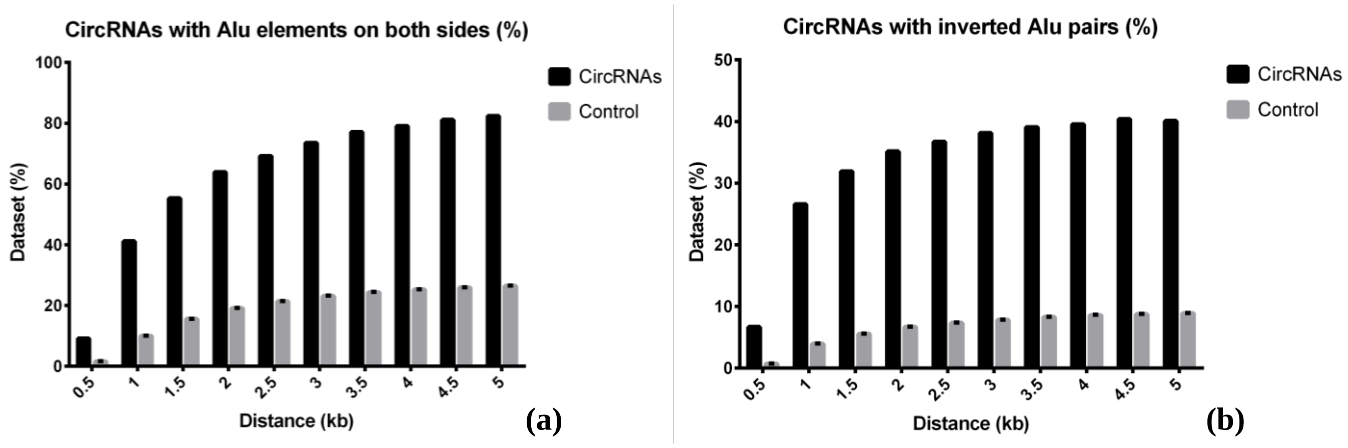


Figure B3. Cumulative number of circRNAs which have Alu elements on both flanks **(a)** and circRNAs which have available Alu elements to form at least 1 inverted pair between flanks **(b)**. In both plots, circRNAs are significantly different from control for all distances (****, $P < 10^{-4}$).

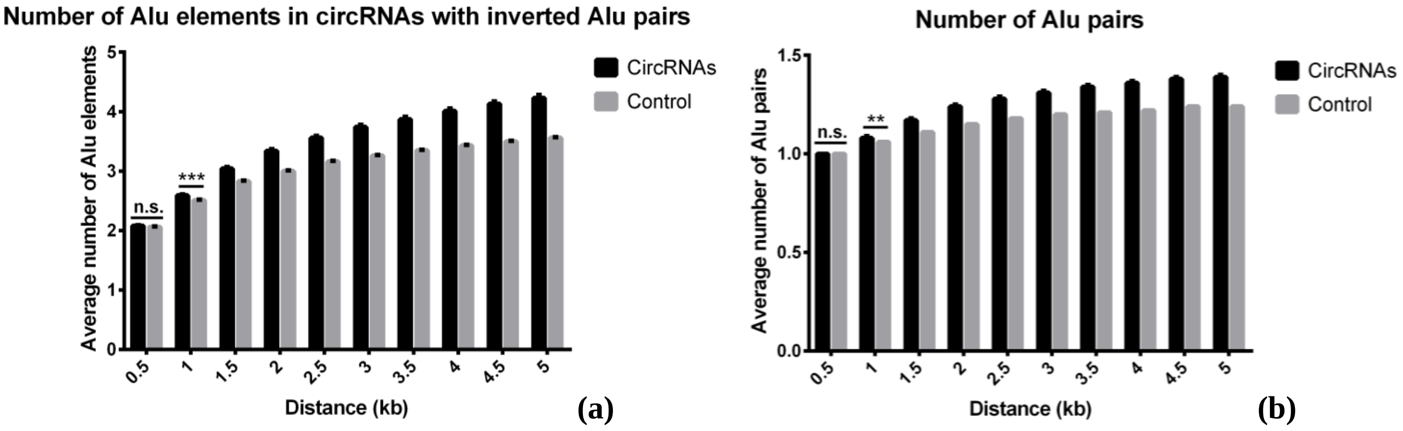


Figure B4. Cumulative average number of available Alu elements in circRNAs which have at least 1 Alu pair **(a)** and predicted number of formed Alu pairs **(b)**. In both plots, plotted p-values: (***), P < 0.001; (**), P < 0.01; (*n.s.*), P > 0.05. Unplotted p-values represent highly significant differences (****, $P < 10^{-4}$).

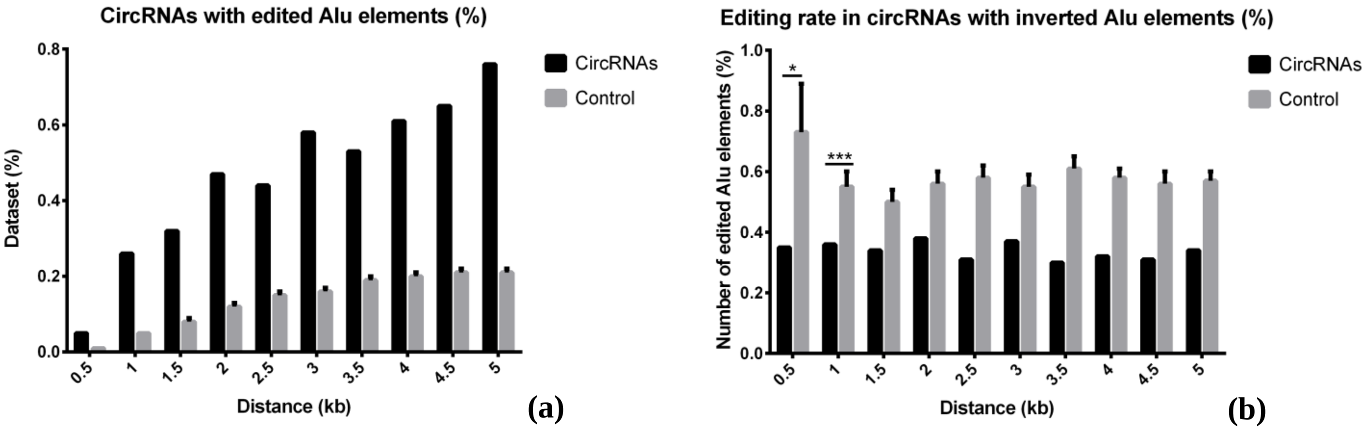


Figure B5. Cumulative number of circRNAs which have edited Alu elements **(a)** and average editing rate in Alu elements flanking circRNAs **(b)**. In both plots, unplotted p-values represent highly significant differences (****, $P < 10^{-4}$). In **(b)**, plotted p-values: (***), P < 0.001; (*), P < 0.05.

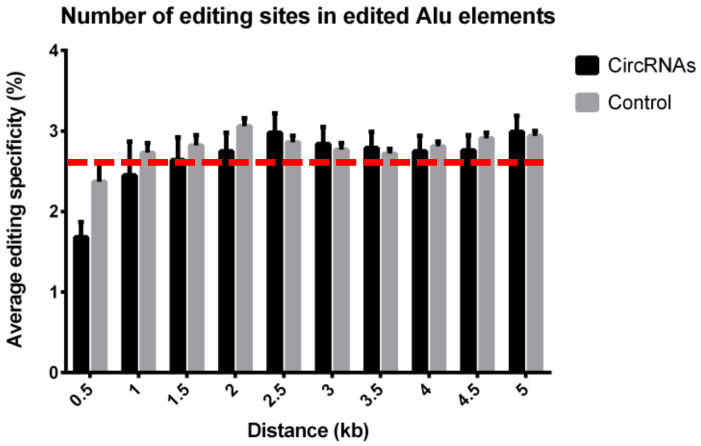


Figure B6. Average editing specificity in edited Alu elements. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05). Dashed horizontal line represent a possible plateau basal level of editing specificity.

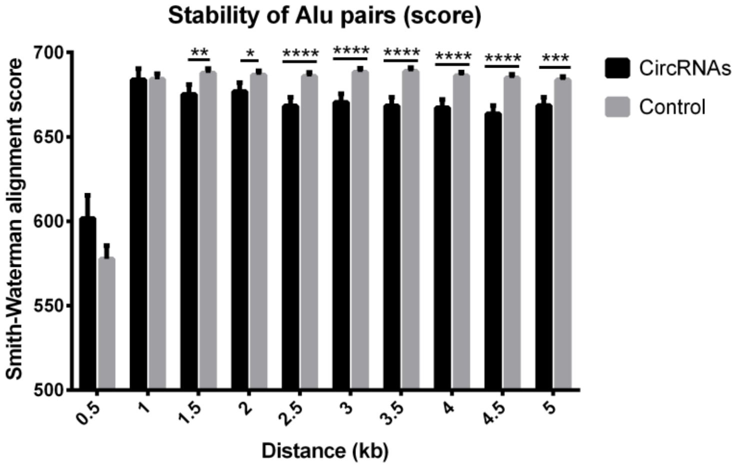


Figure B7. Average stability of Alu pairs. Plotted p-values: (****), $P < 10^{-4}$; (***), $P< 10^{-3}$; (**), $P < 0.01$; (*), $P < 0.05$. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05).

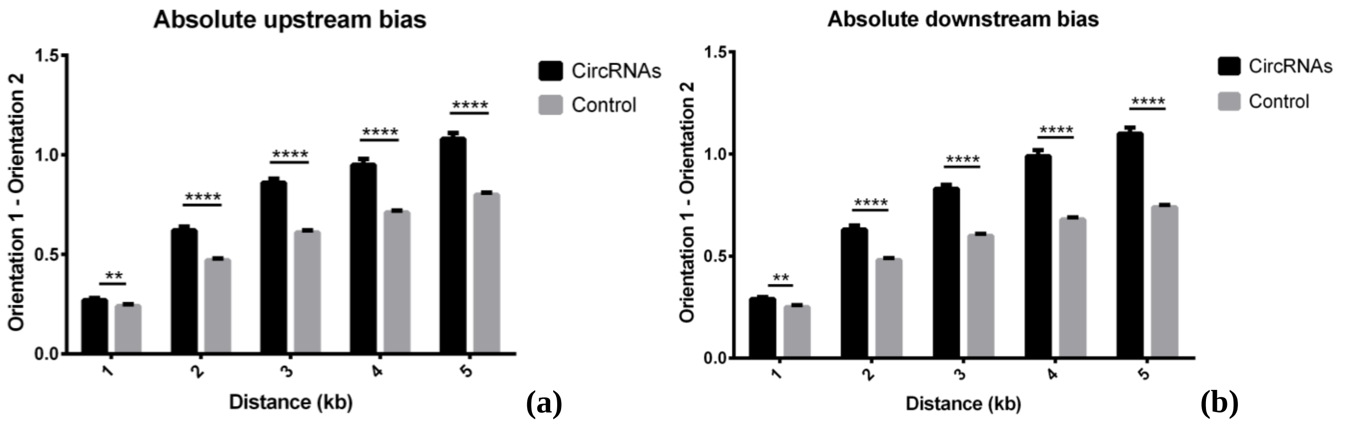


Figure B8. Average absolute orientation bias in upstream **(a)** and downstream **(b)** flanks. Plotted p-values: (****), $P < 10^{-4}$; (**), $P < 10^{-2}$. Unplotted p-values represent insignificant differences (*n.s.*, $P > 0.05$).

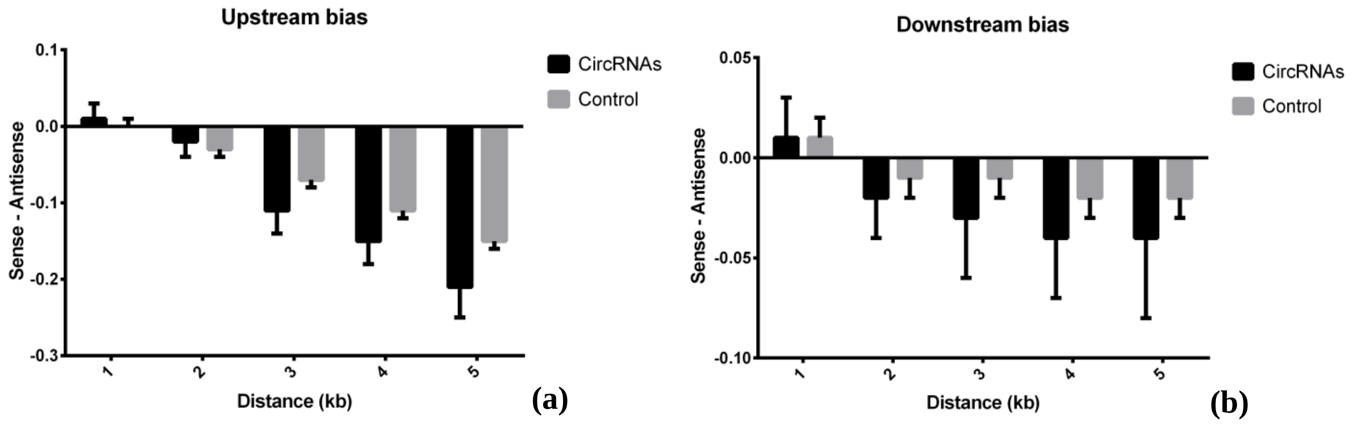


Figure B9. Average orientation bias in in upstream **(a)** and downstream **(b)** flanks. In both plots, circRNAs do not differ significantly from control for all distances (*n.s.*, $P > 0.05$).

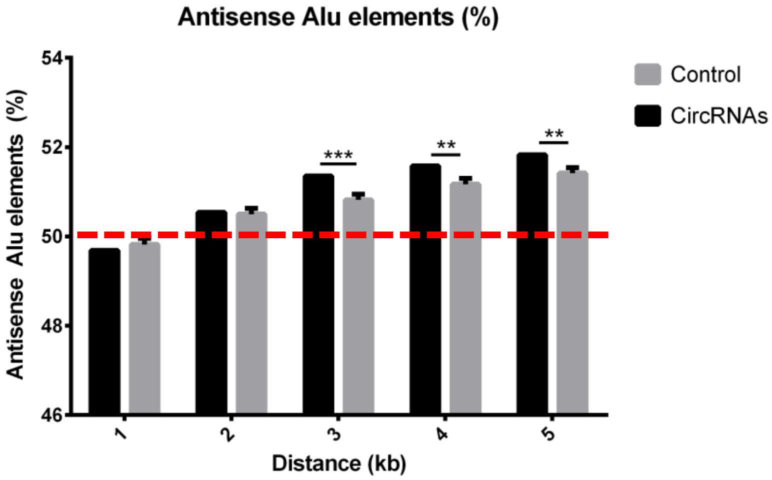


Figure B10. Abundance of Alu elements in antisense orientation. Plotted p-values: (***), *P* < 10^{-3}; (**), *P* < 10^{-2}. Unplotted p-values represent insignificant differences (*n.s.*, *P* > 0.05). Dashed horizontal line represents an overall abundance of 50%.