From the comparative study of a circRNA originating from an mammalian ATXN2L intron to understanding the genesis of intron lariat-derived circRNAs

Annie Robic, Chloé Cerutti, Julie Demars, Christa Kühn

▶ **To cite this version:**

Annie Robic, Chloé Cerutti, Julie Demars, Christa Kühn. From the comparative study of a circRNA originating from an mammalian ATXN2L intron to understanding the genesis of intron lariat-derived circRNAs. Biochimica et Biophysica Acta - Gene Regulatory Mechanisms , 2022, 1865 (4), pp.194815. 10.1016/j.bbagrm.2022.194815 . hal-03664418v1

HAL Id: hal-03664418

https://hal.inrae.fr/hal-03664418v1

Submitted on 11 May 2022 (v1), last revised 27 May 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

BBA - Gene Regulatory Mechanisms 1865 (2022) 194815

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

From the comparative study of a circRNA originating from an mammalian *ATXN2L* intron to understanding the genesis of intron lariat-derived circRNAs

Annie Robic [a,*], Chloé Cerutti [a], Julie Demars [a], Christa Kühn [b,c]

[a] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France
[b] Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196 Dummerstorf, Germany
[c] Faculty of Agricultural and Environmental Sciences, University of Rostock, 18059 Rostock, Germany

ARTICLE INFO

Keywords:
Intron lariat
Splicing
Intron excision
ciRNA genesis
Branch point
sisRNA

ABSTRACT

Circular intronic RNAs (ciRNAs) are still unexplored regarding mechanisms for their emergence. We considered the *ATXN2L* intron lariat-derived circular RNA (ciRNA-ATXN2L) as an opportunity to conduct a cross-species examination of ciRNA genesis. To this end, we investigated 207 datasets from 4 tissues and from 13 mammalian species. While in eight species, ciRNA-ATXN2L was never detected, in pigs and rabbits, ciRNA-ATXN2L was expressed in all tissues and sometimes at very high levels. Bovine tissues were an intermediate case and in macaques and cats, only ciRNA-ATXN2L traces were detected. The pattern of ciRNA-ATXN2L restricted to only five species is not related to a particular evolution of intronic sequences. To empower our analysis, we considered 221 additional introns including 80 introns where a lariat-derived ciRNA was previously described. The primary driver of micro-ciRNA genesis (< 155 nt as ciRNA-ATXN2L) appears to be the absence of a canonical "A" (i.e. a "tn**A**" located in the usual branching region) to build the lariat around this adenosine. The balance between available "non canonical-A" (no ciRNA genesis) and "non-A" (ciRNA genesis) for use as a branch point to build the lariat could modify the expression level of ciRNA-ATXN2L. In addition, the rare localization of the 2′-5′ bond in an open RNA secondary structure could also negatively affect the lifetime of ciRNAs (macaque ciRNA-ATXN2L). Our analyses suggest that ciRNA-ATXN2L is likely a functionless splice remnant. This study provides a better understanding of the ciRNAs origin, especially drivers for micro ciRNA genesis.

1. Introduction

Over the past decade, high-throughput sequencing-based methods have revolutionized our knowledge and understanding of gene expression. The study of transcriptomes keeps improving our understanding of the RNA splicing step. The splicing of the primary transcript (pre-mRNA) comprises removal of introns (mainly non-coding sequences) and ligation of exons (mainly coding sequences). The intron excision is mediated by the spliceosome, which recognizes at least three genetic features within each intron: the 5′ splice site (5′SS), the 3′ splice site (3′SS), and the branch point sequence (BPS) [1–3]. In the first step, the 2′-hydroxyl of an internal nucleotide (the "branch point" of the BPS) attacks the phosphodiester upstream of the intronic 5′SS. The activation of the spliceosome closes the distance between the 5′SS and the BPS containing the two reactive groups of the branching reaction. In the second step, the 3′-hydroxyl of the 5′-exon attacks the phosphodiester at the intronic 3′SS, ligating the exons and creating a lariat-intron product. The intron excision leads to the release of a lariat molecule [4], which is a circular RNA (circRNA) produced by branching from the 5′ end (5′SS) of the intron close to the branch point by a 2′-5′ bond and while keeping a 3′ tail [5–7]. Usually, the lariat is only an intermediate molecule that is rapidly degraded by the debranching enzyme and exonucleases [8]. The hypothesis is that some lariats could escape degradation to become intronic circRNA precursors [9,10]. The first lariat-derived circRNAs were observed in 2012 [11] and characterized in 2013 [9]. Among circRNAs containing only intronic sequences, those derived from the lariat are the most numerous but when a circularization event of the entire intron occurs after the main splicing, an intron circle could be observed [12–14].

The branch point region plays a critical role in RNA splicing catalysis

* Corresponding author.
E-mail addresses: annie.robic@inrae.fr (A. Robic), chloe.cerutti@inrae.fr (C. Cerutti), julie.demars@inrae.fr (J. Demars), kuehn@fbn-dummerstorf.de (C. Kühn).

https://doi.org/10.1016/j.bbagrm.2022.194815
Received 3 February 2022; Received in revised form 12 April 2022; Accepted 14 April 2022
Available online 2 May 2022
1874-9399/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

and the involvement of this region in the genesis of lariat-derived circRNA appears obvious [13,15]. The branch point nucleotide is usually an adenosine and in the intron lariat it is always linked 2′–5′ to the 5′ end of the intron [15–18]. The canonical sequence of the branch point region usually consists of five to seven nucleotides and appears somewhat species dependent [13,15–20]. Usually in mammals, the lariat intron is built around an "A" located in a region of 20 nucleotides and this "A" is preceded by two nucleotides (a "T" and "N") [17]. The canonical environment of the "A" acting a branch point to build the lariat can be defined by what we will henceforth call the nucleotide trio "tn**A**". In the vast majority of introns, the branch point nucleotide is located between 18 and 37 bases upstream of the 3′ splice site [13,18,19]. Several studies [12,14,21] showed that most lariat-derived circRNAs originate from intron lariats do not use an adenosine at the branch point but rather a 'C', or more rarely, a 'T' or 'G'. The hypothesis would be that the lariat debranching enzyme does not have its full capacity to hydrolyze the 2′-5′ bond when the branch point nucleotide is not an "A" [22,23]. Nevertheless, a notable part of lariat-derived circRNA including at least 300 nucleotides could be derived from a lariat built with an "A" [9].

Intronic circRNAs belong to a previously described class of RNAs called 'stable intronic sequence RNAs' (sisRNAs; [10,11,21,24–26]). They are defined as being circular lariat products of splicing inefficiently debranched in the nucleus and exported to the cytoplasm via an NXF1/NXT1-dependent mechanism [14,27]. Two studies [10,28] showed that the circular represented the majority form of stable lariats. The two most studied circRNA containing only intronic sequences (ciRNA) are probably the two lariat-derived circRNAs from the *AKRD52* [9,29] and the *Insulin-2* gene [30,31]. Authors interested in the ciRNA from *ANKRD52* focused on a putative function in the nucleus by suggesting that ciRNA could facilitate R-loop formation by maintaining a locally open RNA structure in vitro. Stoll et al [31] explored other regulatory functions by pointing out that the ciRNA from the *Insulin-2* gene (as many of other ciRNA [14]) is also found in the cytoplasm. The third example of the impact of a ciRNA is still more surprising as the spliced circular intron from human C9ORF72, instead of pre-mRNA, serves as the translation template [27]. These three examples showed that cellular functions of ciRNAs still remain largely elusive.

The detection of intronic circRNAs seems underestimated likely due to technology issues [28]. It was suggested that their small size and probably strong secondary structures near the 2′-5′ junction could be major obstacles to their reliable detection [28,32]. In pubertal porcine testes, we were very surprised to note that the circRNA associated with the highest number of reads was the lariat-derived intronic circRNA from *ATXN2L* (ciRNA-ATXN2L) [33]. To our knowledge, this is the only circRNA that has been identified with such a prominent position in the landscape of a circular transcriptome. We thought at first that it was a feature of the porcine testis [33]. Recently, in describing the circular transcriptional landscape in porcine, and bovine testicular, muscular and liver tissues, we noted that out of the 53 bovine and 80 porcine introns identified as able to produce intronic circRNAs, only ciRNA production derived from the *ATXN2L* lariat was conserved across the two species [34]. Furthermore, we observed a quantitative dominance of one particular circRNA in porcine testis: this was again the circRNA derived from the lariat of the *ATXN2L* gene [34].

We already know that porcine ciRNA-ATXN2L has the particularity of being derived from a very small intron (143 bp) and by a lariat branching with a "non-A" nucleotide [33]. The particularly high expression level of ciRNA-ATXN2L at a biologically sensitive time of testis development (puberty) observed in pigs suggested that it may have a specific function, at least in the pubertal pig testis [33]. We know that *ATXN2L* plays a very important role in embryos [35,36]. However, the detection of this ciRNA in porcine liver and muscle raised questions [34], as the role of *ATXN2L* in these tissues seems less obvious. For this study, we take advantage of total RNA datasets publicly available to investigate ciRNA-ATXN2L in 13 mammalian species with the aim of understanding its genesis and perhaps finding elements about its possible function. We found this ciRNA in five species and our investigations have shown that this presence is disconnected from the phylogeny of the intronic sequences. To improve the understanding of ciRNA generation, we started by considering 64 introns of similar size (< 180 bp like the *ATXN2L* intron), of which 18 were able to produce a ciRNA. Because we suspected that intron length was important for categorizing ciRNAs, we analyzed 157 additional introns. Finally, 120 ciRNAs were analyzed and allows us to identify the drivers of the ciRNAs genesis of and in particular those of smallest size.

2. Materials and methods

2.1. Datasets

For the main analysis, 207 datasets (total-RNAseq from males) were retained: 120 datasets from the project PRJEB33381 (10 mammalian species), 6 porcine datasets from PRJNA506525, 6 bovine datasets from PRJNA471564, 18 rabbit datasets from PRJNA475375, 12 porcine datasets from PRJNA720752, 21 goat datasets from PRJCA002010, 12 sheep datasets from PRJNA613135, and the 6 bovine datasets already used in [34] from PRJEB34570. All details of these datasets are given in Appendix-A.

2.2. Identification of ciRNA-ATXN2L

RNA-seq reads were mapped to the genome reference assemblies of the respective species using the rapid splice-aware read mapper Spliced Transcripts Alignment to a Reference (STAR) [37]. We used the gene annotation provided by Ensembl (Suppl. Doc. 1, Table S1) and selected the single-end alignments mode of STAR (STAR-SE) mapping mates of each pair independently. The detection of circRNAs consists of the identification of reads containing a circular junction and the used approach (CD) has previously been described [34]. CD identifies reads containing a circular junction within those reads that STAR calls "chimeric reads" (CR) from the tabular file (chimeric.out.junction) provided by STAR: subsequently, these reads will be called "circular chimeric reads" (CCRs). Only circRNAs characterized by at least five CCRs were retained in the output file provided by CD. This tool also provides a file reporting all statistics of STAR mapping. The source code of CD is available from https://github.com/GenEpi-GenPhySE/circRNA.git.

ATXN2L is annotated in all 13 mammalian species considered in this study (Suppl. Doc. 1, Table S1). We identified ciRNA-ATXN2L by examining the appropriate region in the putative circRNA output list provided by CD. We suggest quantifying the presence of ciRNA-ATXN2L as a number of CCRs by millions of uniquely mapped reads. The expression (E) of ciRNA-ATXN2L was classified in five classes: high, moderate, weak, and very weak expression correspond to an "$E > 3$", "$1 < E < 3$", "$0.1 < E < 1$", and "$0.01 < E < 0.1$", respectively.

To evaluate the number of linear transcripts produced by the *ATXN2L* gene, we suggest considering the number of split reads (SRs) observed at the exon-exon junction concerned by the ciRNA and reported by STAR. We checked on the six datasets of PRJNA506525 and on the three porcine testicular datasets of PRJEB33381 that the evaluation of mRNA by the number of SR and by RSEM led to identical rankings (RSEM performed for [38]).

2.3. Sequences analyses

We chose to name a ciRNA that comprises less than 155 nucleotides (nt): a micro-ciRNA. The threshold of 155 nt was chosen due to restrictions for the RNA secondary structure analyses performed on these micro-ciRNAs (see also below). The large-ciRNAs considered in this study comprise at least 300 nt and less than 1,000 nt. We define a micro-intron as comprising less than 180 bp and a midi-intron between 330

and 1,100 bp. We called mini-introns the introns with a size intermediate between micro- and midi-introns.

Ensembl reference sequences were considered for all loci analyzed in this study [39]. To examine porcine ciRNAs in addition of ciRNA-ATXN2L, we prioritized dataset SRR8237163, containing a lot of porcine ciRNAs previously identified [12,33,34]. For considering additional bovine ciRNAs, we selected bovine introns already known to be concerned by ciRNA [34]. All additional porcine and bovine ciRNAs were previously described in testis. In addition, we considered six ciRNAs originating from four marmoset micro-introns. They were selected from the output files generated in this study. The marmoset micro-ciRNAs retained were detected in brain and in testis. When we examined introns from genes concerned by ciRNA production in pigs, we identified five genes (*NUP188*, *PITRM1*, *ALDOC*, *DNEP*, and *PAF1*) containing several introns with small length. For investigating introns not involved in ciRNA production, we selected introns from these five genes in pigs, cattle and marmosets with the desired alternative spectrum of length (mini, midi and large).

Kadri et al. [17] studied the intronic branch point sequences in bovine introns and proposed the identification of the most probable nucleotide used as branch point by using the prediction software BPP [40]. They published this prediction for 179,476 bovine introns and we screened these results (reported in the table "Supplementary data 1" of [17]) to select the data corresponding to the bovine introns considered here.

To study the RNA secondary structures, the RNAfold web server was used. As this tool [41] was not designed for circular molecules, several presentations of the input sequence were tested. As this tool analyzes the predicted 2D-structure by plotting base-pairing probabilities, we retained as much as possible the structures with a high probability. We retain the structure proposed with the Minimal Free Energy (MFE). Initially, we had planned to perform these analyses with ciRNAs comprising up to 180 nucleotides. We noticed that by using this tool, we were able to determine a structure associated with high probabilities up to 155 nt, but not beyond. Therefore, an experimental threshold of 155 nt was used to define micro-ciRNAs.

For identifying potential small RNA gene features in ciRNA-ATXN2L, we used Rfam [42] and R2DT [43]. As these tools were not designed for circular molecules, we tested several presentations of the input sequence.

Sequences were analyzed to identify binding sites for (human) RBPs (RNA binding proteins) with the beRBP tool. The input sequence was the original sequence after duplication because this tool [44] was not designed for considering circular molecules or small sequences.

A set of sequences were aligned using CLUSTAL W [45] in order to build the phylogenetic tree using SeaView-4 [46]. The BioNJ approach was chosen with observed distances [47]. Default parameters were used. Other alignments were performed with MultAlin [48].

2.4. Statistical analyses

All the statistical analyses were carried out using *R* (v.4.0.2) [49]. Due to the small sample sizes, significant differences between circRNA proportions from contingency tables were identified with Fisher's exact test (*fisher.test* function from *R stats* package v.4.0.2) [49]. Significant differences between theoretical and observed frequency were identified with exact binomial test (*binom.test* function from *R stats* package v.4.0.2) [49]. A p-value less than 0.05 was considered as statistically significant. Hierarchical clustering was performed for the circRNA classification (*hclust* function from R *stats* package v.4.0.2) using the *ward.D2* agglomeration method from a distance matrix [50]. Several distance measures were computed, first with the *Euclidean* method (*dist* function from R *stats* package v.4.0.2), then with the *Ahmad* method adapted for mixed variables (*distmix* function from R *kmed* package v.0.4.0) [51,52]. The hierarchical clustering figures were performed with the function *fviz_dend* from the *R factoextra* package V.1.0.7 [53].

2.5. Web tools

ClustalW: https://www.phylogeny.fr/one_task.cgi?task_type=clustalw
Rfam: https://rfam.xfam.org/search#tabview=tab1
beRBP: http://bioinfo.vanderbilt.edu/beRBP/predict.html
R2DT: https://rnacentral.org/r2dt
RNAfold: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
Sequence logo: https://weblogo.berkeley.edu/logo.cgi
MultAlin: http://multalin.toulouse.inra.fr/multalin/

3. Results and discussion

3.1. Characterization of ciRNA-ATXN2L in 13 mammals

3.1.1. Detection of circRNAs and identification of the ciRNA-ATXN2L

We retained 207 total-RNA-seq datasets publicly available and consequently considered 13 mammalian species (pig, macaque, marmoset, opossum, mouse, rat, dog, horse, cat, rabbit, cattle, goat and sheep). In a first round of 207 analyses, ciRNA-ATXN2L was not detected in any datasets of goat testis, sheep testis, cattle liver and cattle muscle. Accordingly, we tried to improve the detection power of this ciRNA in these species-tissue combinations by merging these 45 datasets by age of animals within species to obtain 12 collective datasets. This strategy increased the power and enabled the detection of ciRNA-ATXN2L in an additional tissue-species combination (bovine liver).

This study was carried out with datasets produced in eight different laboratories, and we are aware that the circRNA analyses are difficult when merging datasets from different origins [34]. Consequently, we searched to obtain quality criteria to validate ciRNA-ATXN2L detection especially for datasets in which ciRNA-ATXN2L seems to be missing. Details of these analyses are reported in Suppl. Doc. 2. Thus, in summary we performed 174 (162+12) initial circRNA detections of which 166 were considered for further exploration leading to the consideration of 44,255 million uniquely mapped reads. (For details, see Suppl. Doc. 1, Table S2).

3.1.2. Expression level of ciRNA-ATXN2L

We are aware that when analyzing datasets from different origins, it is necessary to be very careful when giving and comparing statistics on ciRNA expression. Statistics about the full analysis are reported in Suppl. Doc. 3 (Figure S1) and we chose retaining only semi-quantitative results. The expression of ciRNA-ATXN2L was classified in five classes (not detected, very low, low, moderate, and high) and are reported in Table 1. From these results, we noticed a large variability between datasets of different origins, even though they were from the same tissue and species (porcine testis from PRJNA506525, PRJEB33381 and PRJNA720752). We suggest retaining the highest expression observed for each species. In summary, we will retain that in pigs and rabbits, ciRNA-ATXN2L could be detected at very high levels compared to cats and macaques showing only low ciRNA-ATXN2L expression and cattle with an intermediate level.

The ciRNA-ATXN2L can constitute a large proportion of the circular transcriptome landscape, since we detected nearly 11,000 and 8,200 reads (CCRs) in pig and rabbit datasets, respectively. For six and four pig and rabbit datasets, ciRNA-ATXN2L-specific CCRs represent more than 2% of total CCRs, the highest proportion being 3.2% and 5.20 %, respectively. In the six pig testis datasets from PRJNA506525, ciRNA-ATXN2L is the top-ranked circRNA in terms of CCR number and with a large distance to the second-ranked one. In all three rabbit testis datasets from PRJEB83381, ciRNA-ATXN2L is also the first circRNA ranked in terms of CCR, but this ranking was very tight.

3.1.3. ciRNA-ATXN2L and mammalian species

ATXN2L is annotated in all 13 mammalian species considered in this

Table 1
Screening to identify ciRNA-ATXN2L. The screening of 166 total-RNA-seq datasets publicly available allowed the consideration of four tissues (testis (T), muscle (M), liver (L) and brain (B)) for 10 mammalian species, of three tissues for cattle (T, M, and L) and only testicular tissue for goat and sheep. The number in front of the tissue abbreviation indicates the number of data sets available for this tissue. The expression of ciRNA-ATXN2L was codified in five classes by using the number of CCRs specific to ciRNA-ATXN2L by millions of reads uniquely mapped. High, moderate, weak, and very weak expression (E) correspond to an "E > 3", "1 < E < 3", "0.1 < E < 1", and "0.01<E<0.1", respectively. The highest E was retained for each species concerned by ciRNA-ATXN2L production (in bold characters). All details were reported in Suppl. Doc. 3 (Figure S1).

	Tissues investigated	Number of reads screened	Detection of ciRNA-ATXN2L				
		(millions of reads uniq. mapped)	High	Moderate	Weak	Very weak	Not detected
Cat	3T- 3M-3L-3B	4,579				**1M**	3T, 2M, 3L, 3B
Cattle	6T- 1M-1L	1,892			**5T**	1T, 1L	1M
Dog	3T- 3M-3L-3B	3,744					3T, 3M, 3L, 3B
Goat	7T	2,563					7T
Horse	3T- 3M-3L-3B	5,579					3T, 3M, 3L, 3B
Macaque	3T- 3M-3L-3B	6,003				**1T**	2T, 3M, 3L, 3B
Marmoset	3T- 3M-3L-3B	4,779					3T, 3M, 3L, 3B
Mouse	3T- 2M-2L-3B	1,603					3T, 2M, 2L, 3B
Opossum	3T- 1M-3L-1B	1,236					3T, 1M, 3L, 1B
Pig	21T- 3M-3L-1B	3,910	**6T**	1T, 1M	2T, 2M, 3L	3T, 1B	9T
Rabbit	3T-21M-3L-3B	3,469	**3T, 3M**	2M, 2L, 2B	15M, 1L, 1B	1M	
Rat	3T- 3M-3L-3B	3,980					3T, 3M, 3L, 3B
Sheep	3T	918					3T

study (Suppl. Doc. 1, Table S1) and for comparative sequence studies, we suggest retaining focus on the last part of the its exon-intron structure exon/introns as shown in Figure 1A. The intron concerned by the generation of ciRNA-ATXN2L (intron_C) follows an exon of 207/213 bp. A phylogenetic tree was built considering the 13 sequences from *ATXN2L* intron_A, exon_AC and intron_C, respectively (Figure 1B). We did not observe large differences between these phylogenetic trees. Moreover, there is no clustering of the five species concerned by the production of a ciRNA-ATXN2L in the tree built with the 13 sequences from *ATXN2L* intron_C. We do not observe a special evolution of intron_C sequences that could explain ciRNA-ATXN2L production in five of the species.

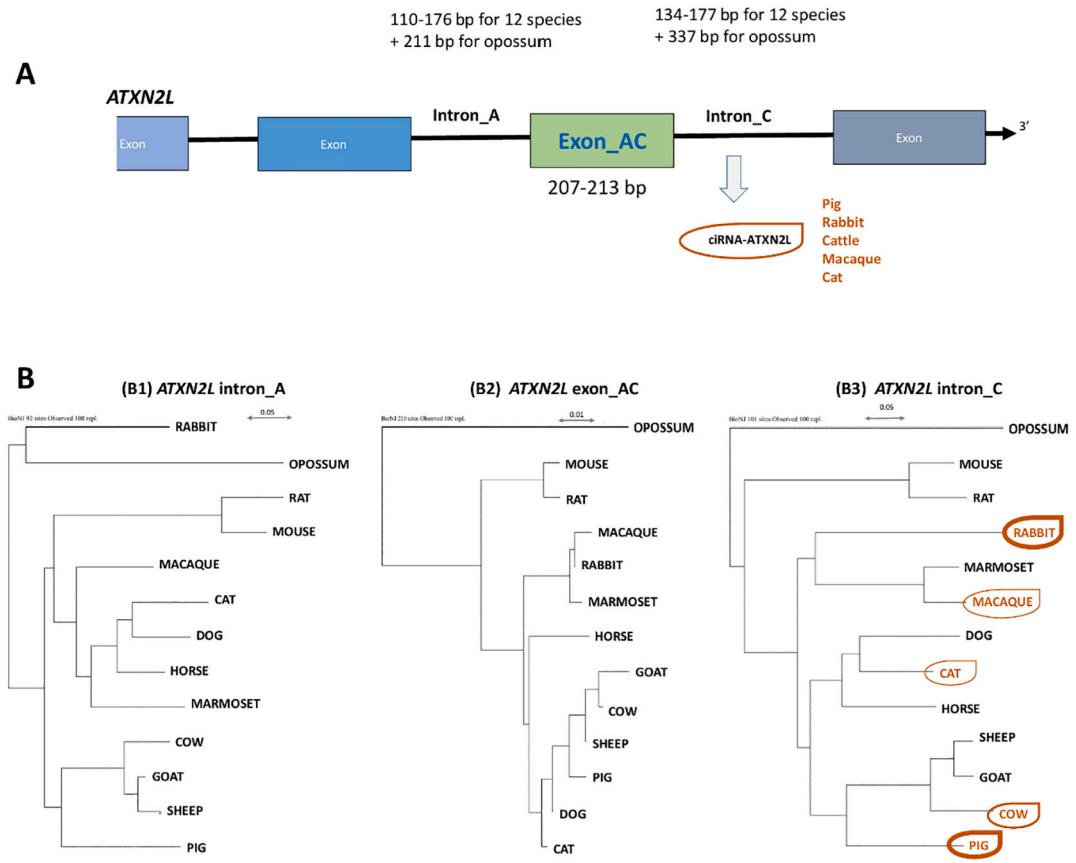

Figure 1. ATXN2L in 13 mammals. (A) Structure of 3′end of *ATXN2L* in mammalian species. The intron_C is the intron concerned by the ciRNA production and is often the last 3′ intron. In the 13 mammalian species, the length of the upstream exon (exon_AC) included between 207 and 213 nucleotides (Table S4). The intron_A, not concerned by ciRNA, is located upstream this exon. The last exon represented is the exon located downstream of the intron_C, which is often the last one and contains a non-coding part of the *ATXN2L* gene. (B) Phylogenetic trees built with the 13 mammalian sequences from *ATXN2L* intron_A (B1), exon_AC (B2) and intron_C (B3). The name of five species concerned by ciRNA-ATXN2L production is indicated in brown characters on A and B parts. Sequences and data are available in Suppl. Doc. 4-6.

3.2. Sequences analyses relative to ciRNA-ATXN2L in 13 mammals

3.2.1. Intronic sequences

We next focused on specific characteristics in the intronic sequences to try to explain why a ciRNA derived from intron_C can be detected in only five of the 13 species and why a ciRNA derived from intron_A was never detected. First, we looked at the sequence composition of the intron_C and intron_A of *ATXN2L*. Indeed it was reported that an intronic sequence with a high level of GC content is important for ciRNA [12,29] or more generally for sisRNA [26] production. For intron_C, among eight species with a high GC content (>54%), we find the five species, where the ciRNA-ATXN2L was observed (pig, rabbit, cow, cat and macaque), in addition to sheep, marmoset and goat. We observed also a high GC content in intron_A sequences of pig, rabbit, cow, cat, macaque, sheep, marmoset, dog, and horse (Suppl. Doc. 7, Table S4).

The analysis of the intronic sequences did not explain why a ciRNA derived from intron_C can be detected in only five of the 13 species and why a ciRNA derived from intron_A was never detected. The two introns are classical introns (5′SS-3′SS = GT-AG) in all species. As the comparative analysis of 5′SS and 3′SS was not informative, we will suggest a careful study of the branching region. (see Section 3.2.2.)

The ciRNA-ATXN2L detection provided the opportunity to examine the individual sequence of the *ATXN2L* intron_C specific to those animals considered by the datasets. We examined the nucleotide sequence of reads mapped on the intron to detect a sequence variation from the reference genome. Sequence variations could lead to other forms of

Figure 2. Searching the nucleotide convenient for branch point use. Usually, the lariat intron is built around an "A" located in a 20 nt region and with an environment defined by the trio sequence motif "tn**A**". Each "A" was indicated in capital letter and each "tn**A**" located in this usual BPS region (i.e. canonical "A") was highlighted in green. (A) Focus on the 20 nucleotides constituting the usual branching region (indicated by a blue background) of the ATXN2L intron_A and intron_C and identification of a canonical "A". (B) Focus on the 3′ end of the intron_C of ATXN2L. (B1) For these six species, no canonical "A" was available. The search for an alternative "A", convenient for branch point use, was made from 3′ to 5′ direction in the intronic sequences. The first "tn**A**" located outside the usual branch point region of 20 nt was highlighted in yellow. Alternative trios of nucleotides able to generate a branching around an "A" were underlined. These alternative "A"s were highlighted in yellow. In pigs and cats, it is highly unlikely that this first "tn**A**" is used to form a lariat around an "A", which would lead to a lariat with a very long 3′tail (69 and 76 nt, respectively). (B2) For these five species, ciRNAs were identified. The analysis of the sequence at the ciRNA circular junction identified the nucleotide used as branch point for lariats leading to the detected ciRNAs. This nucleotide was indicated as bold character and the nucleotide trio was highlighted in cyan.

ciRNA-ATXN2L with possible different characteristics. This analysis was performed for the six pigs included in PRJNA506525, the nine rabbits (German Zika) from PRJNA475375, the six bulls from PRJNA471564, and for the macaque and the cat, including one positive dataset for ciRNA-ATXN2L. The analysis of intronic sequences provided no evidence of a distinct intronic sequence in these animals.

3.2.2. Focus on branch point region

We know that 90% of branch points occur within 18 to 37 nucleotides upstream of the 3′ splice site [13,18,19] , notably inducing the formation of a 3′tail of 17 to 36 nt. We suggest considering this region of 20 nt as the canonical branching region. Usually, the lariat intron is built around an “A” located in this 20 nt region and with an environment defined by the trio “tn**A**” [13,15–20]. Indeed, analysis of the region of *ATXN2L* intron_A sequences comprising the putative branch point identify one or two “tn**A**” in the usual BPS for the 13 species considered (Figure 2A and Suppl. Doc. 5). Analysis of the region of *ATXN2L* intron_C sequences was more complex, as a "tn**A**" was identified for only seven (mouse, opossum, horse, rat, goat, sheep, and dog) of the 13 species considered (Figure 2A and Suppl. Doc. 4). In the absence of a canonical “A” in the BPS of *ATXN2L* intron_C (marmoset, cat, cattle, macaque, rabbit, and pig), alternative trios of nucleotides able to generate a branching around an “A” were searched (Figure 2B1). An alternative “A” can be a “tn**A**” located outside the canonical branching region or an “A” with another nucleotide environment (highlighted in yellow in Figure 2B1). In Figure 2B1, we propose a view of the “A”s selected (*a priori*) for use as branch point.

The presence of ciRNA indicates the inability of the debranching enzyme to hydrolyze the 2′-5′ bond. The analysis of the sequence at the ciRNA circular junction led to the identification of the nucleotide used as branch point. In pigs, four forms of ciRNA-ATXN2L were detected. The circular isoform including 116 nt was the most prevalent form followed by those including 114, 117 and 115 nt (named respectively pig ciRNA-ATXN2L 1/4, 2/4, 3/4 and 4/4). In rabbits, four forms were observed and the form including 121 nt was dominant. In cattle, two forms were observed and the form including 120 nt was dominant. Whatever tissue was considered, the major form was always the same within species and represent at least 80% of CCRs associated with the ciRNA. Only two minor forms of ciRNA-ATXN2L derived from a lariat built around a "T", all others derived from a lariat built around a “C” (Figure 2B2). We noticed the cat ciRNA-ATXN2L and the major form of cattle, pig, and rabbit ciRNA-ATXN2L were generated using a “ct**C**” for the branching (Figure 2B2). No "ct**C**" is found in the usual branching region for the macaque *ATXN2L* intron_C.

When the last 45 intron_C nucleotides of the sheep, goat, and pig *ATXN2L* were aligned, in addition to the fact that the sheep and goat sequences are identical, we observed a block of nucleotides with 26/27 conserved (Figure 3). This block includes BPS in all three species. It is quite curious to note that the 15 nucleotides located upstream and the 11 nucleotides located downstream of the branching point, which is used in pigs and which leads to the major form of ciRNA-ATXN2L are conserved downstream and upstream of the "A", probably used as a branching point, in sheep and goats.

3.3. Analysis of 37 micro-ciRNAs

3.3.1. Addition of 25 micro-ciRNAs outside of ATXN2L

The twelve ciRNA-ATXN2L identified in pig, cattle, cat, macaque and rabbit were very small (114-149 nt), and we proceeded considering other ciRNAs with similar size to empower our analysis. We suggest calling ciRNAs with less than 155 nt nucleotides as micro-ciRNAs and their parent introns as micro-introns. For further analysis, we retained the sequences for the *ATXN2L* intron_Cs from the five species with ciRNA-ATXN2L and their respective 12 ciRNA-ATXN2L. These were complemented by eleven porcine, three bovine and four marmoset micro-introns (105-175 pb) associated with 25 micro-ciRNAs (73-151 nt) (Suppl. Doc. 8 & 9). Among the total of 37micro-ciRNAs considered, 35 derive from a lariat obtained with a “C” at the branch point position and two from a “T”. If we look at the trio of nucleotides present at the branch point, we can note a preferential use of the trio “ct**C**”. This nucleotide trio is the most frequently used among the 37 micro-ciRNAs (12/37) considered, and it is also found in seven major forms of the nine ciRNAs (Suppl. Doc. 10, Figure S2).

3.3.2. Consideration of 76 micro-introns involved or not involved in ciRNA production

For the 5 micro-introns parent of ciRNA-ATXN2L and the further 18 pig/cattle/marmoset micro-introns parent to additional micro-ciRNA, the "A" present in the canonical region of the branch point was a “tn**A**” only for bovine ciRNA-LRP1 (Figure 4, boxes with yellow-green and orange background, for details, see Suppl. Doc. 4 & 8). For the *ATXN2L* intron_A not concerned by the production of ciRNA, one or two “tn**A**” were identified in this region for all of the initially considered 12 species (the *ATXN2L* intron_A of opossum is not a micro-intron). When we add the seven *ATXN2L* intron_C sequences for those species where no ciRNA was detected (the *ATXN2L* intron_C of opossum is not a micro-intron), at least one “tn**A**” was detected in the usual branching region for 18/19 of *ATXN2L* micro-introns not concerned by ciRNA (Figure 4, boxes with white or purple background). To complement our analysis, 10 porcine, 10 cattle and 14 marmoset micro-introns not concerned by ciRNA were selected (Suppl. Doc. 11-13). A “tn**A**” was detected in the usual branching region for these 26/34 micro-introns (Figure 4, boxes with cyan background).

For 45/76 micro-introns we identified at least one “A”, probably very convenient and attractive for branch point use. This “A” has a canonical environment (“tn**A**”) [13,15–20] and the 3′-tail of the lariat generated by the link 2′-5′ is short. The difference in the availability of a “tn**A**” in the usual branching region is statistically significant between micro-introns concerned (1/23) and not concerned (44/53) by ciRNA production (p-value = 5.364e-11, Suppl. Doc. 25). The comparison of the 3′ end of intron_C of *ATXN2L* in sheep, goat with that of pigs showed that a single loss of an “A” included in a “tn**A**” could be crucial and lead to genesis of ciRNA (Figure 3). This criterion of presence or absence of “tn**A**” in the BPS appears as essential for the detection of a micro-ciRNA. The absence of a "tn**A**" in the canonical branching region would be the first driver of micro-ciRNA formation (Figure 4).

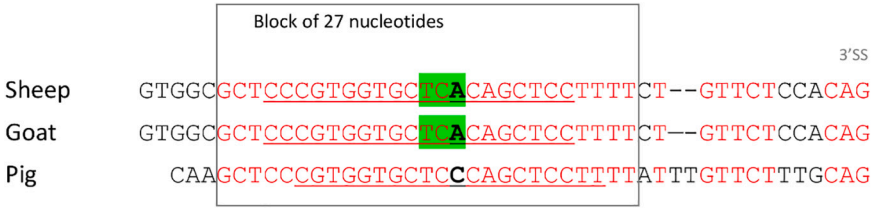

Figure 3. Alignment of the 45 nucleotides located at the 3′end of the intron_C of ATXN2L in goat, sheep and pig. Aligned nucleotides were indicated in red color. The 20 nucleotides included in the BPS were underlined. In sheep, goat sequences, the “A” probably used as a branch point and its canonical environment (“tnA”) are highlighted in green. The nucleotide “C”, which is used in pigs and which leads to the major form of ciRNA-ATXN2L was indicated in bold character.

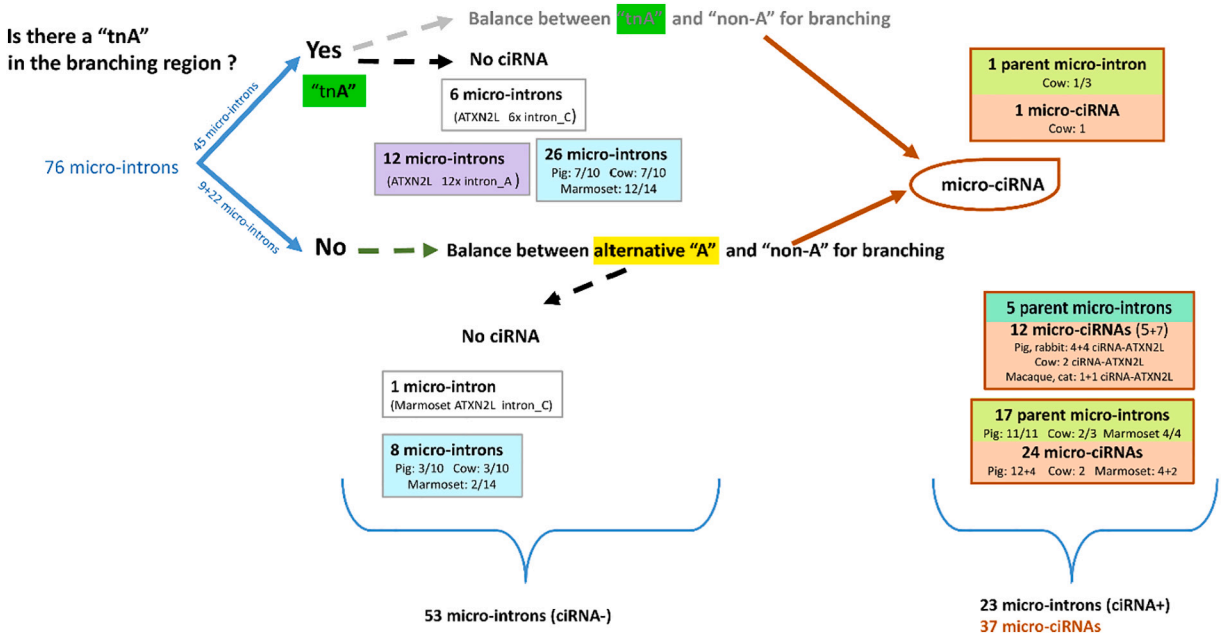

Figure 4. The absence of a "tnA" is the first driver of micro-ciRNA genesis: the difference in the availability of a "tnA" in the usual branching region is statistically significant between micro-introns concerned (1/23 introns ciRNA-) and not concerned (44/53 introns ciRNA+ , which are represented in boxes framed in brown) by ciRNA production. The two boxes with no background color refer to the *ATXN2L* intron_C from eight species not concerned by ciRNA production and the box with green background groups the *ATXN2L* intron_C from the other five species. The box with purple background groups the *ATXN2L* intron_A from 12 species. The two boxes with yellow-green background refers to 18 non *ATXN2L* micro-introns ciRNA+. The two boxes with blue background refer to 34 non *ATXN2L* micro-introns ciRNA-. Concerning ciRNAs, 37 micro-ciRNAs were analyzed and they were presented with a brown background. For details, see Suppl. Doc. 4, 5, 8-13.

3.3.3. Probable competition at the branch point to generate ciRNA-ATXN2L

All 37 micro-ciRNAs considered were derived from lariats built with a "non-A" located in the canonical branching region. We can hypothesize that when not all intronic drivers are present for classical excision (associated with rapid destruction of the intronic lariat), the excision of the intron could lead to a ciRNA. We noticed that the "ct**C**" trio was very often used for the excision of introns leading to a micro-ciRNA and that the resulting ciRNAs were most often the dominant form when there were more than one ciRNA produced. We suggest that whether or not a "ct**C**" is available for branching may explain some of the observed differences in the efficiency of ciRNA-ATXN2L production. Among the five species where a ciRNA-ATXN2L was observed, the macaque is the only one to have a canonical branching region of this intron containing no "ct**C**". Nevertheless, six micro-introns used another trio of nucleotides rather than using the "ct**C**" present in the usual branching region. This study emphasizes that the nature and environment of the "non-A" available is certainly important for use as a branching point when the canonical branching region does not contain "tn**A**".

Sequence analyses of the 37 micro-ciRNAs and their 23 parent micro-introns (*ATXN2L* intron_C from 5 species, 11 porcine, 3 bovine, and 4 marmoset, further ciRNA+ micro-introns) showed that in the absence of a canonical "A" ("tn**A**") in the canonical branching region, the recruitment of an alternative nucleotide takes place primarily in the branching region. For the marmoset, because no "tn**A**" was present in the canonical branching region of *ATXN2L* intron_C and nevertheless no ciRNA was observed, it would be interesting to increase the number of datasets screened to be sure that indeed a ciRNA-ATXN2L is not present. Moreover, we would like to point out that the only "A" available in the canonical branching region of the *ATXN2L* intron_C of the marmoset is a "ct**A**". This trio of nucleotides was never observed in the BPS region for micro-introns (0/20), which are parent to micro-ciRNA and have at least one "A" not "tn**A**" (Suppl. Doc. 8 & 4). This particularity might explain why no ciRNA-ATXN2L has been detected in marmoset tissues.

The excision of the intron_C of *ATXN2L* for marmoset is very informative and shows that the absence of "tn**A**" in the usual branching region does not necessarily lead to the use of a "non-A" as a branch point. Even when a ciRNA is detected, we cannot state that an alternative "A" (a "tn**A**" located outside the canonical branching region or an "A" with another environment) is never used. For five species (macaque, rabbit, cat, bovine and pig), a competition seems possible between "A" and "C" as branch point and the competition terms could affect the amounts of ciRNA-ATXN2L formed. This type of competition based on the relevance of the nucleotide used as branch point was already described [54]. We would underline that this study does not bring any direct information on the use of the non-canonical "A" but we can accumulate some interesting observations. We noticed that the cat is the only species in our data set to have the opportunity to use a "gg**A**" located in the canonical branching region to build a lariat around an "A" (Figure 2B1). Moreover, the trio "gg**A**" is rarely observed in this region for parent micro-introns having at least one "A" (2/20) concerned by the production of micro-ciRNA. The trio "gg**A**" would be more attractive that the trio "cc**A**" found in bovine, cat, and porcine usual branching regions to build a lariat around an "A". The cattle canonical branching region contains two alternative "A"s ("cc**A**" and "gc**A**") in addition of a distant "tn**A**" (Figure 2B1). It is possible that in cats, the two available non-canonical "A" offer better opportunities to form an *ATXN2L* lariat around an "A" than in cattle. In rabbits and pigs, there is probably only one opportunity to build a lariat around an "A": a "tn**A**" located just outside the canonical branching region for rabbits and a "cc**A**" for pigs. In resume, we suspect that in cats, and cattle, available non-canonical "A" offer better opportunities to form an *ATXN2L* lariat around an "A" than in rabbits or pigs. For these four species, these alternative "A"s are in competition with a "ct**C**" to form a lariat around this branch point. In macaque, opportunities to form an *ATXN2L* lariat around an "A" are intermediate (a distant "tn**A**" and a distant "gg**A**") but the second competitive nucleotide to use as a branch point is a "C" in a trio "tt**C**". In resume, the genesis of the cat ciRNA-ATXN2L, and to a lesser extent in cattle, appears to be less favored than in pigs and rabbits.

3.3.4. Analysis of RNA secondary structure at the circular junction of micro-ciRNAs

We investigated predicted features of RNA secondary structures of ciRNA-ATXN2L from different species. These *in silico* analyses are relevant to propose modeled secondary structures for small RNAs, even circular [55], probably because they are not able to bind to RBPs. To describe different parts of the RNA secondary structure proposed by the tool RNAfold web server [41], we referred to the types defined by [56]. We tested several presentations of the sequence (see Suppl. Doc. 14) to explore the potential secondary structure of micro-ciRNAs. More than the secondary structure of the entire circular RNA, we were particularly interested in the position of the circular junction (2′-5′ bond) within a region where a secondary structure was predicted by the tool RNAfold with high probabilities.

This search was performed for all ciRNA-ATXN2L observed in the five species. We noted some difficulties with the macaque to obtain a proposition with good probabilities (Suppl. Doc 14, Figure S4). For the majority of ciRNA-ATXN2L, the nucleotide used as a branch point was located in the closing pair of a hairpin loop (Figure 5). The ciRNA-ATXN2L from macaque ciRNA-ATXN2L is distinguished from other ciRNA-ATXN2L by a positioning of the 2′-5′ junction within an open RNA structure. Similar analyses were performed on the further porcine, bovine and marmoset micro-ciRNAs considered in this study. Examination of the predicted structures for these 25 micro-ciRNAs showed that the circular junction was never located in an open RNA structure (Figure 5). Because of the small size of ciRNA-ATXN2L, it is not possible to improve the accuracy of the computational modeling of the secondary structure by integrating experimental information [57].

Unfortunately, we cannot study the secondary structures of the lariat itself due to its inherent degradation, and ciRNA is the only trace of the branching. We have no proof that secondary structures observed in the region of the circular junction of a ciRNA are the source of the inability of the debranching enzyme to hydrolyze the 2′-5′ bond resulting from a branching around a non-A. Moreover, when we compare two ciRNAs from the same intron, one very hypothetical from a "tnA" branching and the other observed and resulting from a “tgC” branching, we do not observe any major difference in structure (cattle LRP1- Suppl. Doc. 14, Figure S5). The example with the differences observed in the secondary structure of macaque ciRNA compared to other ciRNAs can be interpreted that the macaque ciRNA-ATXN2L is more fragile than other ciRNAs.

3.3.5. ciRNA-ATXN2L in mammals: expression and “conservation”

As shown above, the genesis of the cat ciRNA-ATXN2L, and to a lesser extent in cattle, appears to be less favored than in pigs and rabbits. The balance between available "alternative-A" (no ciRNA genesis) and "non-A" (ciRNA genesis) in this region for use as a branch point could determine the level of ciRNA-ATXN2L present. It is somewhat difficult to classify the conditions of macaque ciRNA-ATXN2L genesis compared to other species. Nevertheless, we showed that in the macaque ciRNA-ATXN2L the 2′-5′ bond would be less protected by RNA secondary structures than in pig, cat, cattle, or rabbit ciRNA-ATXN2L. This configuration does not seem to have a secondary impact of using a “tt**C**” as BPS, since porcine ciRNAs from *TCN2*, *PAF1* are also built around a “tt**C**” and do not present this lack of a protective niche. We are convinced that the secondary structure present at the 2′-5′ junction can affect the lifetime of the macaque ciRNA-ATXN2L.

Talhouarne and Gall [14] identified seven ciRNAs deriving from orthologous human-mouse introns. Stoll et al. [31] showed that a ciRNA from the *Insulin-2* gene exists in humans, mice and rats. The current study provides a report on the conservation of a ciRNA across five mammalian species but disconnected to the phylogeny of these species. We know that ciRNAs are much rarer than exonic circRNAs for which orthologous conservation studies have been more numerous [34,58–61]. The ability to produce exonic circRNA seems to be mainly linked to the genomic structure of genes (topic reviewed by [62,63] in 2019, [12,60,64,65]), which is often conserved. We showed that the “conservation” of ciRNA-ATXN2L in mammals is related to intronic sequence features.

3.4. Search for a possible function for ciRNA-ATXN2L

The ciRNA-ATXN2L comprises only 116-121 nucleotides for three species (pig, rabbit, and cattle), and 138 and 149 nt for cats and macaques, respectively, which corresponds to the linear intronic sequence. For identifying potential small RNA gene features in ciRNA-ATXN2L, we propose to conduct a structural analysis. We integrated in our analysis the hypothesis that its function could be obtained by its circular structure. The main form of ciRNA-ATXN2L characterized in cattle, pig, and in rabbit was analyzed with the main objective to include the sequence spanning the circular junction. These query sequences did not match any Rfam families and any structures of the R2DT templates. As the folding of a given RNA depends not only on the sequence information but also

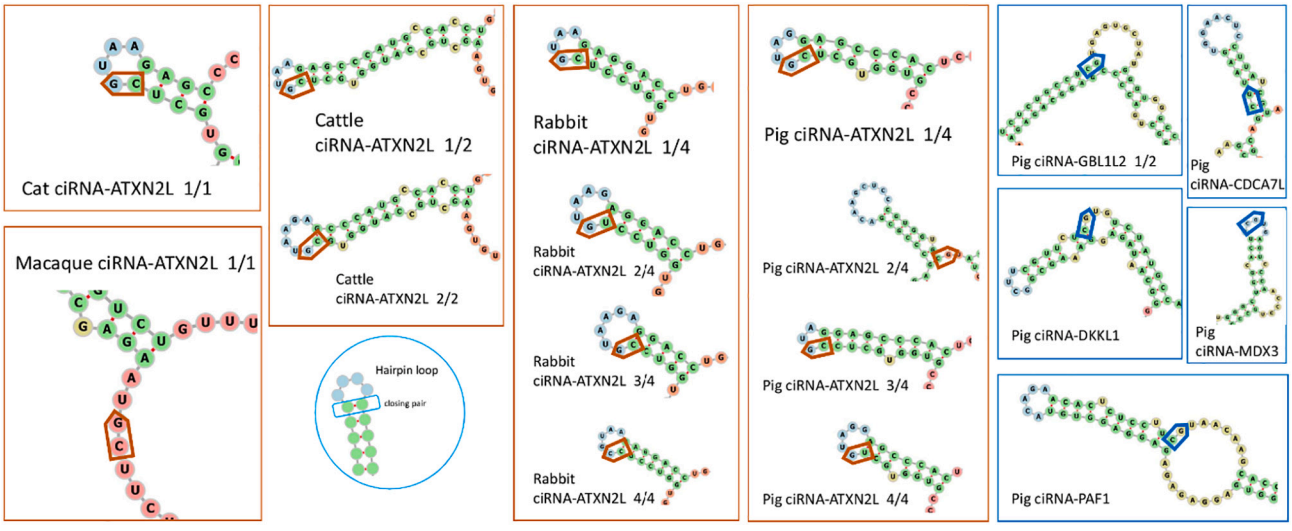

Figure 5. RNA secondary structures at the 2′-5′ bond region proposed by RNAfold. Nucleotides involved in a hairpin loop are indicated in blue and those involved in a stem are indicated in green by the tool (RNAfold, “forna” format). The 12 ciRNA-ATXN2L were presented in brown frameworks. For cattle, rabbit and pig, the different ciRNAs are presented in the order of highest frequency. Five porcine ciRNA were presented in blue frameworks. We added an arrow to indicate the position of the 2′-5′ bond.

on the RNA environment [66], the porcine, rabbit and bovine ciRNA-ATXN2L were analyzed to identify binding sites for RBPs [44]. Even though all human RBPs were considered, no result was obtained.

Even though this study was performed with datasets from multiple origins, it was still possible to compare the number of linear transcripts with those of circular intronic transcripts between datasets simultaneously produced. Considering the six porcine testis samples from PRJNA506525, we observed a ratio ciRNA/mRNA between 4 to 26 for *ATXN2L*, namely ciRNA-ATXN2L is more abundant than *ATXN2L* linear transcripts (all details were reported in Suppl. Doc. 15, Figure S6). In rabbit muscle (PRJNA720752), the ratio ciRNA/mRNA varied from 0.54 to 32 (Chinese Qixin) and from 0.82 to 20.5 (German Zika). When we considered only datasets from rabbit muscle from German Zika animals, we noted the ratio ciRNA/mRNA increased with age (Table S6).

According to Li et al. [29], RNAse H1 is able to degrade ciRNA involved in a competition with the pre-mRNA to form an R-loop with the DNA. Pursuit of this hypothesis would lead us to suggest that ciRNAs with no function in transcription regulation are the least easily degraded. However, among the three porcine testicular datasets from PRJEB33381, the one with highest ciRNA-ATXN2L expression was also the one with the highest expression for the gene *RNASEH1* (RSEM evaluation performed for [38]). We reported similar observations for the rabbit muscle: these datasets presenting the highest expression of *RNASEH1* (linear transcript evaluated by SRs) had the highest ratio ciRNA/mRNA of *ATXN2L* (Suppl. Doc. 15, Table S6). The ratio ciRNA/mRNA appeared to increase with age in rabbit muscle. This is similar to the accumulation of exonic circRNA in neurons with age [67]. In contrast to exonic circRNA, where we can assume a possible competition between two ways to mature the pre-mRNA (splicing and back splicing generating respectively mRNA and exonic circRNA) [68,69], ciRNA is only a secondary product of the splicing. As it has already been reported that introns can be precursors of small non-coding RNA [70,71], several analyses were performed to identify a possible precursor inside sequences of ciRNAs, but without success (see above). The ciRNAs are produced in the nucleus by a pathway that could be described as passive, are exported to the cytoplasm by a non-dedicated mechanism [14] and appear accumulating with age in muscle despite the increase of the transcriptional activity of *RNASEH1*. Moreover, it is highly likely that ciRNA-ATXN2L is unable to interact specifically with targeted RBPs or miRNA. We also showed that the presence of this ciRNA was not associated with a particular evolution of the intronic sequences concerned. These characteristics suggest that ciRNA-ATXN2L probably has no function either in the nucleus or in the cytoplasm.

3.5. Additional analyses: differences in ciRNA genesis related to the size of the parent intron?

The first investigations reported above were made in the context of the excision of micro-introns and the formation of micro-ciRNAs. Several studies have reported ciRNAs [9,10,12,14,21] and we were the first to underline the existence of these micro-ciRNAs [33,34]. Our results are in agreement with previous studies reporting [12,14,21] that most ciRNAs originate from intron lariats not using an "A" as a branch point. Nevertheless, Zhang et al. [9] reported a large part of ciRNAs built around an “A”. As we suspect that this difference is due to the selection on size made by Zhang et al. [9], we propose to consider new ciRNAs/introns.

3.5.1. Consideration of additional larger introns/ciRNAs

In the study of Zhang et al. [9], a selection of ciRNAs including at least 300 nt was applied and consequently, we considered at first large-ciRNAs (300-1,000 nt) from midi-introns (330-1,100 bp). We retained twenty porcine midi-introns associated with 25 large-ciRNAs and six bovine midi-introns associated with seven large-ciRNAs. Among these 32 large-ciRNAs, 29 derived from a lariat obtained with a “C” at the branch point position, one from a “T” and two from an “A” (Table 2). The two ciRNAs that derive from a lariat built around an "A" are the respective unique form of porcine ciRNA-SF3B2 and bovine ciRNA-HNRNPU (for details, see Suppl. Doc. 16-18). No difference from the porcine reference sequence was detected in the sequence of the porcine intron of *SF3B2* (dataset SRR8237163) and the bovine intron of *HNRNPU* (dataset SRR775528-530). For the intron branching from *SF3B2* for which a large-ciRNA was identified, we could affirm that at least a part of the lariats was built using an “A” in the trio “cc**A**”. At least a part of lariats of *HNRNPU* was built using the “A” located in position -15 (i.e. downstream the usual branching region) in the trio “aa**A**”. Among these 26 midi-introns, three porcine and two bovine midi-introns contained a “tn**A**” in the usual branching regions (Table 2). Among these five introns, the branching region of the bovine *HNRNPU* intron is very particular since it contains ten "T" and seven "A". No difference from the reference sequence was detected in the sequence of usual branching regions for these five introns. To conclude, we identified only two ciRNAs among the 32 large-ciRNAs built around an "A" while Zhang et al. reported 8/20 [9].

Table 2
Contingency table of introns and ciRNAs according to the branching region characteristics. The availability of a "tn**A**" is statistically different within the group of 234 introns (and within the subgroups of 76 micro-introns, 71 mini-introns, 87 midi-introns) according to their involvement or not in ciRNA production (p-value < 0.001).

	Introns producing no ciRNA			149 introns
	44/53	27/35	53/61	
	Introns producing ciRNAs			85 introns
"tnA" available in the branching region	1/23	8/36	5/26	
	76 micro-introns (90 < bp < 180)	71 mini-introns (181 < bp < 320)	87 midi-introns (330 < bp < 1,100)	234 introns
	micro-ciRNAs	intermediate ciRNAs	large-ciRNAs	
Nucleotide used as a branch point not located in the usual branching region	0/37	0/51	1/32	
"A" is the nucleotide used as a branch point	0/37	1/51	2/32	
	37 micro-ciRNAs	51 ciRNAs	32 large-ciRNAs	120 ciRNAs

Finally, parent introns of/and intermediated size ciRNAs were analyzed (Suppl. Doc. 19-20). We retained 28 porcine mini-introns associated with 41 ciRNAs and 8 bovine midi-introns associated with 10 ciRNAs (Table 2). Among these 51 ciRNAs, 45 derived from a lariat obtained with a “C” at the branch point position, three from a “T”, two from a “G” and one from an “A”. We noticed that bovine ciRNA-SIN3B derives from a lariat built around a "tn**A**". No difference from the bovine reference sequence was detected in the concerned intron sequence (dataset SRR775527). Among these 36 mini-introns, eight mini-introns contained a “tn**A**” in the usual branching regions (Table 2).

To improve the understanding of the ciRNAs genesis, the branching regions of 34 other mini- and 60 midi-introns not concerned by ciRNA were analyzed. Adding the information from the midi opossum *ATXN2L* intron_C, at least one “tn**A**” was detected in the usual branching region for 53 midi-introns of 61 analyzed as introns not concerned by ciRNA (Table 2, Suppl. Doc. 4 & 21-23). Adding the information from the mini opossum *ATXN2L* intron_A, at least one “tn**A**” was detected in the usual branching region for 27 mini-introns of 35 analyzed as introns not concerned by ciRNA (Table 2, Suppl. Doc. 5 & 24).

3.5.2. Global comparisons of the 234 introns

As the ciRNAs considered here are only very rarely built around an "A", the first examination will concern the "A" available in the usual branch point region. The total absence of an "A" (Table 3) appears as a statistically significant difference between introns concerned (9/85) and not concerned (2/149) by ciRNA production (p-value = 0.002212, see Suppl. Doc. 25). Nevertheless, this feature concerns a very small number of introns. The availability of a "tn**A**" in the usual branching region appears as a statistically significant difference between introns concerned (14/85) and not concerned (124/149, Tables 2 & 3) by ciRNA production (p-value < 2.2e-16, see Suppl. Doc. 25). This difference is significant for micro-introns (already reported in Section 3.3.2.), mini-introns and midi-introns considered separately (see Suppl. Doc. 25). Fourteen introns with lariat-derived ciRNA contain a "tn**A**" in the BPS and only one is a micro-intron. The availability of a "tn**A**" in introns not involved in ciRNA production (Table 2) is not statistically different between two groups of introns of different size. We know that the mammalian branch point consensus sequence includes a pyrimidine just downstream the "A" acting a branch point [16,17]. The proportion of "tn**A**y" among the "tn**A**n" present in the BPS does not differ significantly depending on whether the intron is involved in ciRNA production or not (9/14 and 71/124). Having more than one "tn**A**" available in the BPS is also not significantly different between introns concerned or not by ciRNAs production and having already one "tn**A**" (1/14 and 42/124 respectively, p-value = 0.0638, see Suppl. Doc. 25).

Analyses of the micro-introns involved in the production of ciRNA-ATXN2L, led us to suspect that some alternative "A "s ("ct**A**" and "gg**A**") would be a better alternative than others to compensate for the absence of a canonical "A" for the branching of the intron lariat around an "A" . We noted no significant difference in frequency of occurrence of trios "ct**A**", "cc**A**", "gc**A**", and "gg**A**" in BPS between ciRNA-affected and non-affected introns (Table 3, and for statistical analyses, see Suppl. Doc. 25). The trio "ct**A**" (6/85) is represented at the expected frequency (1/12), which is not the case for the trios "cc**A**" (57/85), "gc**A**" (32/85) and "gg**A**" (16/85) (Table 3 and Suppl. Doc. 25). When we examined the BPS region for the 59 bovine introns considered here, we noted that 18 introns had no "tn**A**" (6/ciRNA+ and 12/ciRNA-) available in the BPS to perform a canonical branching but had at least one other "A". Now, we suggest looking at the data published by Kadri et al. [17] concerning the prediction of branch point for bovine introns (see Section 2.3.) and we focused on the predictions reported for these 18 bovine introns without "tn**A**". For eight introns (3/ciRNA+ and 5/ciRNA-), the branching is predicted around an "A" located in the usual branching region. It is quite curious to note that each of the trios "cc**A**" and "ct**A**", which are present at extreme frequencies, is proposed by this analysis in three cases. All these observations reinforce our opinion that the "ct**A**" trio must have good abilities to serve as a branch point in the absence of "tn**A**". We reported this complete comparative analysis of the 59 bovine introns in Suppl. Doc. 26.

Except one marmoset micro-intron of *PITRM1* (ciRNA+), one mini-intron (porcine *PAF1,* ciRNA-), and one midi-intron (porcine *CLTB,* ciRNA+), all introns considered in this study are classical introns, the 5′SS is GT and the 3′SS is AG (Suppl. Doc. 4-5, 8, 11-13, 16-17, 19, 21-24). We can conclude that 231 of the 234 considered are classical introns (GT-AG) and could be excised by the U2 way [72]. The possible production of a ciRNA is not related to this data.

Table 3
Analyze of "A" available in the branching region for 234 introns. For analyses, we focused on the 85 introns (23 ciRNA- and 62 ciRNA+) containing no "tn**A**" in their BPS but at least one "A".

149 introns not concerned by ciRNA (ciRNA-)			85 introns concerned by ciRNA (ciRNA+)		
Presence in the BPS	N. of introns concerned		Presence in the BPS	N. of introns concerned	
at least one "tn**A**"	124		at least one "tn**A**"	14	
none "A"	2		none "A"	9	
at least one "A"	23		at least one "A"	62	
	"ct**A**"	3		"ct**A**"	3
	"gg**A**"	4		"gg**A**"	11
	"gc**A**"	11		"gc**A**"	21
	"cc**A**"	16		"cc**A**"	41

3.5.3. Global comparisons of the 120 ciRNAs

In this study, 37 micro-ciRNAs (73-151 nt), 32 large-ciRNAs (313-990 nt), and 51 with an intermediate size were considered (Table 2). In summary, 120 ciRNAs originating from 85 introns (and 84 genes) were selected, thus 35 correspond to a minor form of another ciRNA within the same intron. On the other hand, the three ciRNAs derived from a lariat built around an "A" are major forms. Among the 120 ciRNAs considered, porcine ciRNAs are in the majority in our data set with 86 units. The size of these 120 ciRNAs goes from 73 to 992 nt with a median at 205 nt. When we sort them according to their size (Suppl. Doc. 27, Table S10), the first ciRNA that derives from an intron with a "tn**A**" in the BPS is ranked at the position #8, but the "tn**A**" is not used to build the ciRNA lariat (cow, 93 nt). The three ciRNAs built around an "A" are ranked at position #78 (cow, 259 nt), #93 (pig, 319 nt), and #108 (cow, 457 nt), and only the first was built around a canonical "A".

3.5.4. Analysis of nucleotide trio used for the branching of the lariat with existing derived ciRNA

We then analyzed the frequency of nucleotide trios observed at the circular junction related to the two extreme classes of ciRNAs: 37 micro-ciRNAs and 32 large-ciRNAs. Twenty nucleotide trios were identified, and to simplify this analysis, nucleotide trios were classified in ten categories. Both types of ciRNAs favor the same five trios, but the order of preferential use appears to differ (Figure 6). Nevertheless, statistical analyses did not reveal significant differences (p-value = 0.55, Suppl. Doc. 25). Eighteen ciRNAs are derived from introns while a canonical "A" was available in the BPS and only one was built around this "tn**A**". We did not observe a significant difference in the use of trios "ct**C**" (3/18) or "tg**C**" (6/18) between these 18 and the 102 others (24/102 and 26/102 respectively).

In addition, we suggest again looking at the data published by Kadri et al. [17] to analyze the branch points predicted for the 18 bovine ciRNA+ introns (22 ciRNAs) investigated in our study (Suppl. Doc. 26). Among these 18 introns, the prediction retained the "tn**A**" in the usual branching region each time it existed, i.e. 4 times. Only one of these "tn**A**" is involved in the excision of the corresponding intron with production of ciRNA (bovine *Sin3B*). Among the twelve introns having a BPS including at least one alternative "A", only for three introns the nucleotide suggested as branch point is the "C involved in the excision of the corresponding intron with production of a ciRNA (bovine ciRNA-VAC14, ciRNA-CHD5, and ciRNA-ATXN2L 2/2). Finally, only for 4/18 introns we observed a match between the "predicted branch point" nucleotide and the nucleotide involved in ciRNA production (three "C" and one "A"). These analyses reinforce the idea that the nucleotide involved in the circular junction of a ciRNA is probably not the only one used as a branch point for the excision of the considered intron.

3.5.5. Clustering of 120 ciRNAs

The classification proposed above is only arbitrary and is only justified by technical considerations (separation between micro and intermediate ciRNAs: because we observed that structural analyses by *RNAfold* demand a maximum size of 155 nt for reliable output) or by historical considerations (>300 nt as Zhang et al. [9]). We conducted a global analysis of the 120 ciRNAs selected to delimit these two groups of extreme size ciRNAs. Three parameters (ciRNA size, simplified branch point (nucleotide trio divided in ten categories), type of "A" present or absent in the BPS, see Suppl. Doc.27 Table S10) were considered to perform clustering. At first, we performed a classical hierarchical clustering of these data using *Euclidean* distances. We reported a simplified

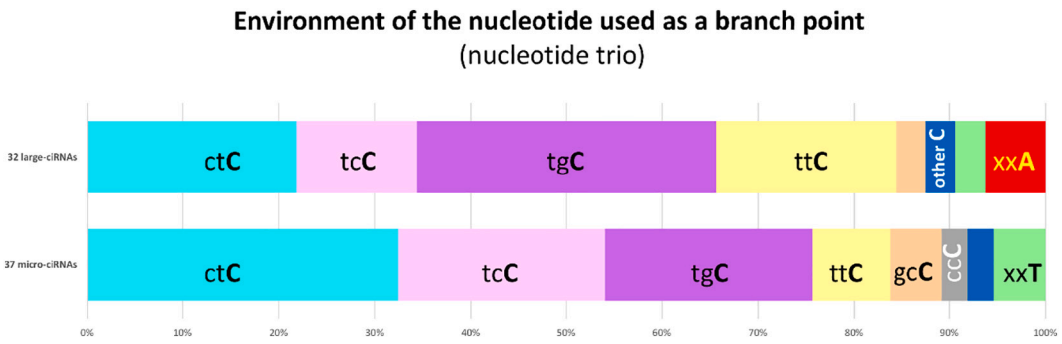

Figure 6. Use frequency of the different nucleotide trio present at the circular junction. In the trio, the nucleotide used as a branch point is written in capital letter. Ten categories of nucleotide trio were considered but among the 32 large- and 37-micro-ciRNA selected, only nine were represented.

version of the dendrogram generated (the original is available in Suppl. Doc. 27, Figure S7A) in the Figure 7A to highlight essential points. The dendrogram shows the splitting into two groups (#1-96 in brown and #97-120 in blue on Figure 7A) and seven subgroups #1-42; #43-72; #73-96; #97-105; #112-117, #118, and #119 & 120). The dendrograms presented in Figure 7B-C were obtained using a distance matrix computed with the *Ahmad* method (the original is available in Suppl. Doc. 27, Figures S7B & 7C respectively). They showed the existence of two groups (in pink and in cyan, Figure 7B) and four groups (in red, green, yellow and purple in Figure 7C), respectively. In fact, these three approaches led to the identification of two major groups #1-96 and #97-120 (Figure 7A: brown and blue, 7B: pink and cyan, 7C: red+green and yellow+purple), or seven sub-groups, and all are identical. The size criterion seems to be a determining factor in these clusterings since these seven sub-groups proposed by the three methodologies are groups of size. The additional splits of certain sub-groups of ciRNAs highlighted by clustering using Euclidean distances (Figure 7A) are not confirmed in the same terms by clustering using Ahmad distances (Suppl. Doc. 27, Figure S7B and S7C). We conclude that these approaches allowed the clustering of the 96 smallest ciRNAs into three groups of size: 73-166 nt (#1-42); 174-227 nt (#43-72); 235-344 nt (#73-96).

3.5.6. *Examination of sequences at the circular junction*

Zhang et al. reported conserved motifs around the circular junction of ciRNAs. Thus, finally, we examined if we could find respective sequences present at the circular junctions of our datasets. As Zhang et al. [9], we considered the eleven nucleotides upstream the branch point and the seven nucleotides near the 5′SS to propose a consensus sequence for four sets of ciRNAs (Figure 8). The first set consists of all porcine ciRNAs with very small (size <166 nt corresponding to the group ciRNA

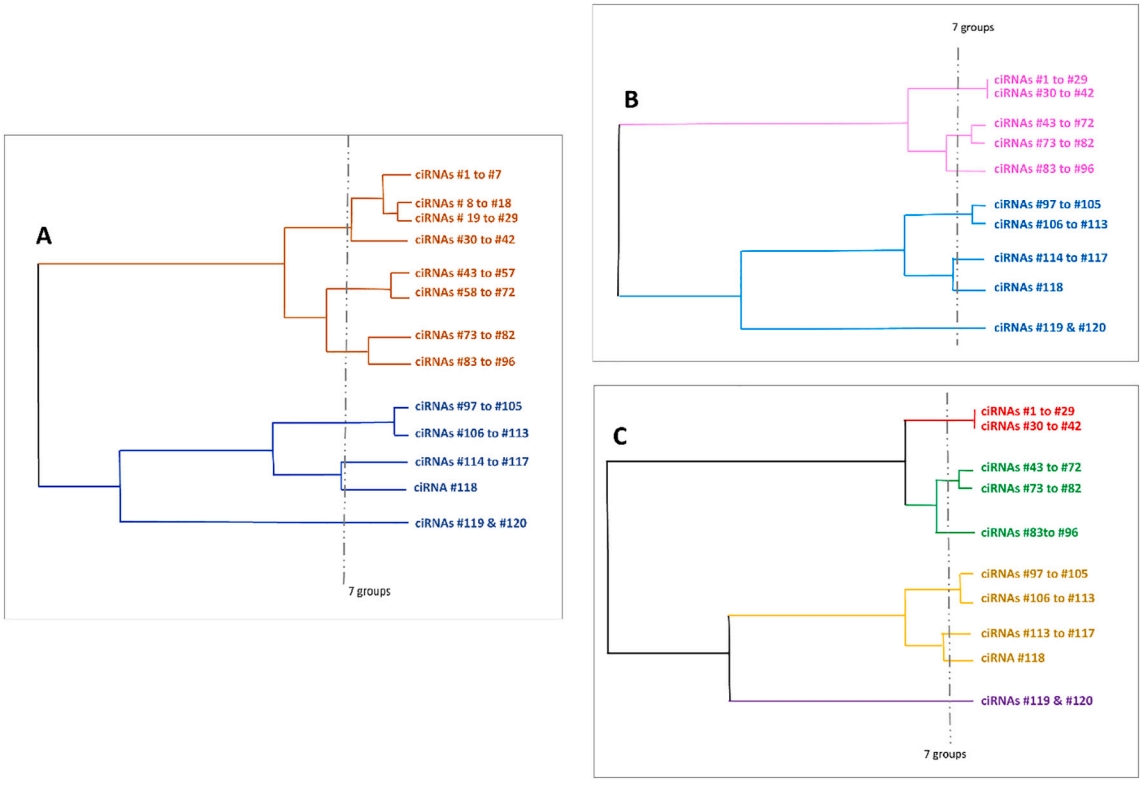

Figure 7. Hierarchical clustering of ciRNAs grouped by size (nt), branch point nucleotide trio and type of "A" present or absence in BPS: The dendrogram of hierarchical cluster analysis based on distance matrix computed by *Euclidean* method separated into two groups (brown and blue, in A). The dendrograms based on distance matrix computed with the *Ahmad* method separated into two groups (pink and cyan, in B) and four groups (red, green, yellow and purple, in C). Each ciRNA was named by its position on the 120 ciRNAs size chart. The three original dendrograms with all 120 ciRNAs are included in Suppl. Doc. 27, in figures S7A, S7B and S7C respectively.

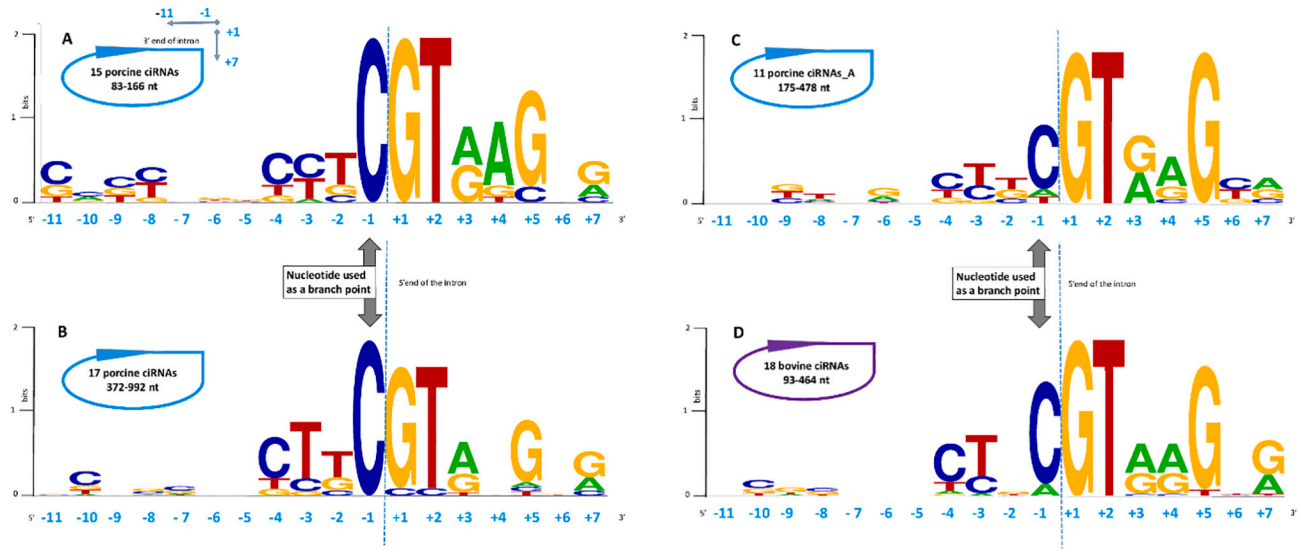

Figure 8. Consensus sequences at the circular junction of ciRNAs. (A) Consensus sequence for porcine micro-ciRNAs (15 ciRNAs with size <166 nt, only major forms were considered). (B) Consensus sequence for porcine large-ciRNAs (17 ciRNAs: 372 < size (nt) < 992). (C) Consensus sequence for porcine ciRNAs having a "tn**A**" in the BPS (but mostly not used for branching) or using an "A" for the lariat branching (11 ciRNAs: 175 < size (nt) < 478) (D) Consensus sequence for bovine-ciRNAs (18 ciRNAs). The eleven nucleotides upstream the branch point and the seven nucleotides near the 5′SS were considered to build these logos. The nucleotide used for the branching (with the position -1) was indicated by a large narrow. For this analysis, only the major forms have been retained.

#1 to #42 defined by clustering approaches, Figure 7) and the second of all porcine ciRNAs with large size (size >372 nt corresponding to the group ciRNA #97 to #120 defined by clustering approaches, Figure 7). In both pattern for these two groups, we did not find the motif "C-rich" of eleven nucleotides nor that "GT-rich" of seven nucleotides suggested previously (Figures 8A & 8B; [9]). For the sequence from the 5′SS, we note a content rather rich in purines and for the nucleotides from the BPS, the content is richer in pyrimidine. These both characteristics appeared more pronounced in porcine micro-ciRNAs (Figures 8A, B). As these two sets did not contain any particular ciRNAs, we investigated a third set constituted by all porcine ciRNAs built around an "A" or having a "tn**A**" in the BPS of the originating intron (Figure 8C). In comparison to the two other porcine consensus sequences, we noticed the presence of a "G" at the position +5, i.e. in the 5′SS. The fourth set of ciRNAs consists of the 18 bovine ciRNAs and does not lead to a logo very different from porcine logos (Figure 8D). Our data demonstrate that there is no conserved motif around the circular junction regardless of species or ciRNA size category within species. These four consensus sequences were constructed with ciRNAs detected in the testes and the tissue criterion could be important. The consensus sequence proposed by Zhang et al. [9] was based on large ciRNAs expressed in mouse RPC cells. Talhouarne and Gall [14] had already observed differences between mouse and human RPCs [14]. In addition, we would like to point out that no RNase R treatment to discard linear RNA was used to generate the datasets considered here. We only analyzed "natural ciRNAs" and we hoped no transient lariat. Indeed, we believe (like Jeck et al. [59]) that the use of RNase R in the preparation of RNA for a sequencing library could transform some lariats into ciRNA.

4. Conclusion

Until now, the drivers of intron excision have been most often studied by considering a large number of introns, however, restricted to a single species. This study shows that cross-species comparison can provide valuable information for understanding intron excision, especially when not all intronic drivers are present for classical excision (associated with rapid destruction of the intronic lariat). This current study has furthered the knowledge and understanding of intron excision and the eventual genesis of ciRNAs with a focus on those we called micro-introns (<180 bp) and micro-ciRNAs (<155 nt). We have suggested that when the lariat of a micro-intron can be built with an "A" sufficiently attractive to constitute a branch point, no micro-ciRNA will be generated. Towards the example of ciRNA-ATXN2L, we showed that the balance between available alternative "A" (no ciRNA genesis) and "non-A " (ciRNA genesis) could determine the level of ciRNA resulting from the excision of micro-introns. In the light of this study, the genesis of ciRNAs seems quite clear, with far fewer exceptions than that of large-ciRNAs. For the future, it would be interesting to investigate more precisely large-ciRNAs including more than 300 nucleotides and built with an "A" as branch point. It was reported that ciRNA-ins2 [30] and ciRNA-C9ORF72 [27] have these both characteristics. Their long RNA sequences probably give them more opportunities to bind to RBPs inducing folding of the RNA and thus could influence the choice of the nucleotide for the branching and provide a protective niche for the circular junction.

The ciRNA-ATXN2L was found in the transcriptome of five mammalian species. The ciRNA-ATXN2L may occupy a major position in the circular transcriptome as in rabbits and pig testis, or be not detected as in eight mammalian species. The data and analyses presented here suggest several explanations for these variations. Nevertheless, no indication that ciRNA-ATXN2L has relevant functions after its genesis were found. The knowledge of the genesis of this ciRNA seems sufficient to explain why it is found in only five of the investigated species. After this study, we think that it probably has no role, a pure remnant of splicing.

We believe that the ciRNA landscape we observe in a tissue results from the conjunction of (at least) three types of factors: (1) The transcription level of the parent gene; (2) The branching of the lariat for intron excision; (3) The lifetime of the ciRNA. In light of this study, branching of the intron lariat appears to be the key splicing step for ciRNA genesis. Moreover, this study suggests that when a ciRNA is observed, there has been competition for branching of an A (most often non-canonical) and a non-A. The level of ciRNA genesis depends on this competition (in addition of the transcription level of the parent gene). This study also suggests that the secondary structures of the RNA may have an impact on the lifetime of the ciRNA once it has been generated.

Credit authorship contribution statement

Annie Robic: Conceptualization, Investigation, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. Chloe Cerutti: Investigation, Software, Statistical analyses, Writing - review & editing. Julie Demars: Conceptualization, Formal analysis, Writing - review & editing. Christa Kühn: Conceptualization, Methodology, Writing - review & editing. All authors have read and agreed to the submitted version of the manuscript.

Funding

This work received no external funding. Nevertheless, studies around circular RNAs were supported by INRAE (GenPhySE and Animal Genetics division) and by the FBN (Institute of Genome Biology).

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

We are grateful to the Genotoul/bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for computing and storage resources. We wish specially to thank Patrice Dehais and Sarah Maman from SIGENAE group.

Appendix A

We retained 207 datasets (total-RNAseq from males). These 207 total-RNA-seq datasets allowed the consideration of 13 mammalian species (pig, macaque, marmoset, opossum, mouse, rat, dog, horse, cat, rabbit, cattle, goat and sheep). We considered four tissues: datasets for testis are available for all 13 species, skeletal muscle and liver for 11 species, and brain for 10 species.

1. The 120 datasets (10 species X 4 tissues X 3 individuals) from the project PRJEB33381 available on ENA [73]. Reads: 2X150bp, Ribo-Zero. The four tissues are Testis, Brain, Muscle and Liver. The ten mammalian species are cat, dog, horse, macaque, marmoset, mouse, opossum, pig, rabbit, and rat.
2. The 6+1 datasets (testis from 7 pubertal pigs) from the project PRJNA506525 on SRA [74]. Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X100bp and 2X125bp; RiboMinus): [12].
3. The 6 datasets (testis from 3 neonatal and from 3 pubertal bulls) from the project PRJNA471564 available on SRA [74]. Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X150bp; Ribo-Zero): [75].
4. The 18 rabbit datasets from muscle (3 ages X 2 breeds X 3 individuals) from the project PRJNA475375 available on SRA [74]. Reference articles including the protocol of the preparation of sequencing libraries (Reads: 2X150bp; Ribo-Zero): [76,77].
5. The 12 porcine datasets from testis (2 ages X 2 breeds X 3 individuals) from the project PRJNA720752 available on SRA [74]. Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X150bp; Ribo-Zero): [78].
6. The 21 datasets from goat testis (7 ages X 3 individuals) from the project PRJCA002010 available on GRA [79]. Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X150bp; Ribo-Zero): [80].
7. The 12 datasets from sheep testis (3 ages X 4 individuals) from the project PRJNA613135 deposited on SRA [74]. Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X150bp; Ribo-Zero): [81].
8. For bovine liver and bovine muscle, we retained the 6 datasets already used in [34] from males (available on ENA [73] under the project PRJEB34570). Reference article including the protocol of the preparation of sequencing libraries (Reads: 2X100bp; Ribo-Zero): [82].

For the main analysis, we chose to not consider the dataset SRR8237163 (ssc_testis_1 in [28] and pig-31 in [25,30]) from the project PRJNA506525. Nevertheless, we used this dataset known for its particular high circRNA content to propose a complementary study on eight ciRNAs.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2022.194815.

References

[1] C.L. Will, R. Luhrmann, Spliceosome structure and function, Cold Spring Harb. Perspect. Biol. 3 (7:a003707) (2011).
[2] M.C. Wahl, C.L. Will, R. Luhrmann, The spliceosome: design principles of a dynamic RNP machine, Cell 136 (4) (2009) 701–718.
[3] M.E. Wilkinson, C. Charenton, K. Nagai, RNA splicing by the spliceosome, Annu. Rev. Biochem. 89 (2020) 359–388.
[4] R.A. Padgett, M.M. Konarska, P.J. Grabowski, et al., Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors, Science 225 (4665) (1984) 898–903.
[5] R. Yoshimoto, N. Kataoka, K. Okawa, et al., Isolation and characterization of post-splicing lariat-intron complexes, Nucleic Acids Res. 37 (3) (2009) 891–902.
[6] W. Keller, The RNA lariat: a new ring to the splicing of mRNA precursors, Cell 39 (3 Pt 2) (1984) 423–425.
[7] S. Borao, J. Ayte, S. Hummer, Evolution of the early spliceosomal complex-from constitutive to regulated splicing, Int. J. Mol. Sci. 22 (22) (2021), 12444.
[8] A. Mohanta, K. Chakrabarti, Dbr1 functions in mRNA processing, intron turnover and human diseases, Biochimie 180 (2021) 134–142.
[9] Y. Zhang, X.O. Zhang, T. Chen, et al., Circular intronic long noncoding RNAs, Mol. Cell 51 (6) (2013) 792–806.
[10] G.J. Talhouarne, J.G. Gall, Lariat intronic RNAs in the cytoplasm of Xenopus tropicalis oocytes, RNA 20 (9) (2014) 1476–1487.
[11] E.J. Gardner, Z.F. Nizami, C.C. Talbot Jr., et al., Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of Xenopus tropicalis, Genes Dev. 26 (22) (2012) 2550–2559.
[12] A. Robic, T. Faraut, S. Djebali, et al., Analysis of pig transcriptomes suggests a global regulation mechanism enabling temporary bursts of circular RNAs, RNA Biol. 16 (9) (2019) 1190–1204.
[13] A.J. Taggart, C.L. Lin, B. Shrestha, et al., Large-scale analysis of branchpoint usage across species and cell lines, Genome Res. 27 (4) (2017) 639–649.
[14] G.J.S. Talhouarne, J.G. Gall, Lariat intronic RNAs in the cytoplasm of vertebrate cells, Proc. Natl. Acad. Sci. U. S. A. 115 (34) (2018) E7970–E7977.
[15] J.M.B. Pineda, R.K. Bradley, Most human introns are recognized via multiple and tissue-specific branchpoints, Genes Dev. 32 (7–8) (2018) 577–591.
[16] K. Gao, A. Masuda, T. Matsuura, et al., Human branch point consensus sequence is yUnAy, Nucleic Acids Res. 36 (7) (2008) 2257–2267.
[17] N.K. Kadri, X.M. Mapel, H. Pausch, The intronic branch point sequence is under strong evolutionary constraint in the bovine and human genome, Commun. Biol. 4 (1) (2021) 1206.
[18] T.R. Mercer, M.B. Clark, S.B. Andersen, et al., Genome-wide discovery of human splicing branchpoints, Genome Res. 25 (2) (2015) 290–303.
[19] A.J. Taggart, A.M. DeSimone, J.S. Shih, et al., Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo, Nat. Struct. Mol. Biol. 19 (7) (2012) 719–721.
[20] D.M. Canson, T. Dumenil, M.T. Parsons, et al., The splicing effect of variants at branchpoint elements in cancer genes, Genet. Med. 24 (2) (2022) 398–409.
[21] H. Saini, A.A. Bicknell, S.R. Eddy, et al., Free circular introns with an unusual branchpoint in neuronal projections, elife (2019), 8:e47809.
[22] A. Jacquier, M. Rosbash, RNA splicing and intron turnover are greatly diminished by a mutant yeast branch point, Proc. Natl. Acad. Sci. U. S. A. 83 (16) (1986) 5835–5839.
[23] A. Kumari, S. Sedehizadeh, J.D. Brook, et al., Differential fates of introns in gene expression due to global alternative splicing, Hum. Genet. 141 (2022) 31–47.
[24] S.N. Chan, J.W. Pek, Stable intronic sequence RNAs (sisRNAs): an expanding universe, Trends Biochem. Sci. 44 (3) (2019) 258–272.
[25] I. Osman, M.L. Tay, J.W. Pek, Stable intronic sequence RNAs (sisRNAs): a new layer of gene regulation, Cell. Mol. Life Sci. 73 (18) (2016) 3507–3519.
[26] J. Jin, X. He, E. Silva, Stable intronic sequence RNAs (sisRNAs) are selected regions in introns with distinct properties, BMC Genomics 21 (1) (2020) 287.
[27] S. Wang, M.J. Latallo, Z. Zhang, et al., Nuclear export and translation of circular repeat-containing intronic RNA in C9ORF72-ALS/FTD, Nat. Commun. 12 (1) (2021) 4908.

[28] A.J. Taggart, W.G. Fairbrother, ShapeShifter: a novel approach for identifying and quantifying stable lariat intronic species in RNAseq data, Quant. Biol. 6 (3) (2018) 267–274.
[29] X. Li, J.L. Zhang, Y.N. Lei, et al., Linking circular intronic RNA degradation and function in transcription by RNase H1, Sci. China Life Sci. 64 (2) (2021) 1795–1809.
[30] D. Das, A. Das, M. Sahu, et al., Identification and characterization of circular intronic RNAs derived from insulin gene, Int. J. Mol. Sci. 21 (12) (2020) 4302.
[31] L. Stoll, A. Rodriguez-Trejo, C. Guay, et al., A circular RNA generated from an intron of the insulin gene controls insulin secretion, Nat. Commun. 11 (1) (2020) 5611.
[32] A. Robic, C. Kühn, Beyond Back splicing, a still poorly explored world: non-canonical circular RNAs, Genes 11 (9) (2020) 1111.
[33] A. Robic, J. Demars, C. Kühn, In-depth analysis reveals production of circular RNAs from non-coding sequences, Cells 9 (8) (2020) 1806.
[34] A. Robic, C. Cerutti, C. Kühn, et al., Comparative analysis of the circular transcriptome in muscle, liver and testis in three livestock species, Front. Genet. 12 (2021), 665153.
[35] N. Gross, M.G. Strillacci, F. Penagaricano, et al., Characterization and functional roles of paternal RNAs in 2–4 cell bovine embryos, Sci. Rep. 9 (1) (2019) 20347.
[36] J. Key, P.N. Harter, N.E. Sen, et al., Mid-gestation lethality of Atxn2l-ablated mice, Int. J. Mol. Sci. 21 (14) (2020), 5124.
[37] A. Dobin, C.A. Davis, F. Schlesinger, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2013) 15–21.
[38] A. Robic, T. Faraut, K. Feve, et al., Correlation networks provide new insights into the architecture of testicular steroid pathways in pigs, Genes 12 (4) (2021), 551.
[39] Ensembl.https://www.ensembl.org/index.html.
[40] Q. Zhang, X. Fan, Y. Wang, et al., BPP: a sequence-based algorithm for branch point prediction, Bioinformatics 33 (20) (2017) 3166–3172.
[41] A.R. Gruber, R. Lorenz, S.H. Bernhart, et al., The Vienna RNA websuite, Nucleic Acids Res 36 (Web Server issue) (2008) W70–W74.
[42] I. Kalvari, E.P. Nawrocki, J. Argasinska, et al., Non-coding RNA analysis using the rfam database, Curr. Protoc. Bioinformatics 62 (1) (2018), e51.
[43] B.A. Sweeney, D. Hoksza, E.P. Nawrocki, et al., R2DT is a framework for predicting and visualising RNA secondary structure using templates, Nat. Commun. 12 (1) (2021) 3494.
[44] H. Yu, J. Wang, Q. Sheng, et al., beRBP: binding estimation for human RNA-binding proteins, Nucleic Acids Res. 47 (5) (2019), e26.
[45] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (22) (1994) 4673–4680.
[46] M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building, Mol. Biol. Evol. 27 (2) (2010) 221–224.
[47] O. Gascuel, BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data, Mol. Biol. Evol. 14 (7) (1997) 685–695.
[48] F. Corpet, Multiple sequence alignment with hierarchical clustering, Nucleic Acids Res. 16 (22) (1988) 10881–10890.
[49] Team Rc, R: A Language and Environment for Statistical Computing, Foundation for Statistical Computing, Vienna, Austria, 2020.
[50] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31 (2014) 274–295.
[51] W Budiaji . kmed: Distance-Based k-Medoids. R package version 0.4.0. http://cran.fhcrc.org/web/packages/kmed/vignettes/kmedoid.html 2021.
[52] A. Ahmad, L. Dey, A K-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2007) 503–527.
[53] A. Kassambara, F. Mundt, actoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. http://www.sthda.com/english/rpkgs/factoextra, 2020.
[54] M.D. Molina-Sanchez, A. Barrientos-Duran, N. Toro, Relevance of the branch point adenosine, coordination loop, and 3' exon binding site for in vivo excision of the sinorhizobium meliloti group II intron RmInt1, J. Biol. Chem. 286 (24) (2011) 21154–21163.
[55] S. Petkovic, S. Graff, N. Feller, et al., Circular versus linear RNA topology: different modes of RNA-RNA interactions in vitro and in human cells, RNA Biol. (2021) 1–10.
[56] P. Danaee, M. Rouches, M. Wiley, et al., bpRNA: large-scale automated annotation and analysis of RNA secondary structure, Nucleic Acids Res. 46 (11) (2018) 5381–5394.
[57] G. De Bisschop, D. Allouche, E. Frezza, et al., Progress toward SHAPE constrained computational prediction of tertiary interactions in RNA structure. Non-coding, RNA 7 (4) (2021) 71.
[58] J.U. Guo, V. Agarwal, H. Guo, et al., Expanded identification and characterization of mammalian circular RNAs, Genome Biol. 15 (7) (2014) 409.
[59] W.R. Jeck, J.A. Sorrentino, K. Wang, et al., Circular RNAs are abundant, conserved, and associated with ALU repeats, RNA 19 (2) (2013) 141–157.
[60] G. Santos-Rodriguez, I. Voineagu, R.J. Weatheritt, Evolutionary dynamics of circular RNAs in primates, elife 10 (2021), e69148.
[61] R. Dong, X.K. Ma, L.L. Chen, et al., Increased complexity of circRNA expression during species evolution, RNA Biol. 14 (8) (2017) 1064–1074.
[62] D.D. Pervouchine, Circular exonic RNAs: when RNA structure meets topology 1862 (11-12) (2019) 194384.
[63] J.R. Welden, S. Stamm, Pre-mRNA structures forming circular RNAs 1862 (11-12) (2019) 194410.
[64] C. Ragan, G.J. Goodall, N.E. Shirokikh, et al., Insights into the biogenesis and potential functions of exonic circular RNA, Sci. Rep. 9 (1) (2019) 2048.
[65] D. Cao, Reverse complementary matches simultaneously promote both back-splicing and exon-skipping, BMC Genomics 22 (1) (2021) 586.
[66] F. Tahi, T.T.V. Du, A. Boucheham, In silico prediction of RNA secondary structure, Methods Mol. Biol. 1543 (2017) 145–168.
[67] H. Gruner, M. Cortes-Lopez, D.A. Cooper, et al., CircRNA accumulation in the aging mouse brain, Sci. Rep. 6 (2016) 38907.
[68] H.M. Li, X.L. Ma, H.G. Li, Intriguing circles: conflicts and controversies in circular RNA research, RNA 10 (9) (2019), e1538.
[69] T. Shao, Y.H. Pan, X.D. Xiong, Circular RNA: an important player with multiple facets to regulate its parental gene expression, Mol. Ther. Nucleic Acids 23 (2021) 369–376.
[70] F. Hube, D. Ulveling, A. Sureau, et al., Short intron-derived ncRNAs, Nucleic Acids Res. 45 (8) (2017) 4768–4781.
[71] G.J.S. Talross, S. Deryusheva, J.G. Gall, Stable lariats bearing a snoRNA (slb-snoRNA) in eukaryotic cells: a level of regulation for guide RNAs, Proc. Natl. Acad. Sci. U. S. A. 118 (45) (2021), e2114156118.
[72] I.V. Poverennaya, M.A. Roytberg, Spliceosomal introns: features, functions, and evolution, Biochemistry. Biokhimiia 85 (7) (2020) 725–734.
[73] ENA. The European Nucleotide Archive is a part of ELIXIR architecture in EMBL-EBI. accessed 2 September 2021;https://www.ebi.ac.uk/ena/browser/home.
[74] SRA. Sequence Reads Archive in NCBI, National Center biotechnologies Information. accessed 2 September 2021;https://ngdc.cncb.ac.cn/gsa/.
[75] Y. Gao, S. Li, Z. Lai, et al., Analysis of long non-coding RNA and mRNA expression profiling in immature and mature bovine (Bos taurus) testes, Front. Genet. 10 (2019) 646.
[76] L. Kuang, M. Lei, C. Li, et al., Identification of long non-coding RNAs related to skeletal muscle development in two rabbit breeds with different growth rate, Int. J. Mol. Sci. 19 (7) (2018), 2046.
[77] X. Zhang, C. Zhang, C. Yang, et al., Circular RNA, microRNA and protein profiles of the Longissimus dorsi of Germany ZIKA and Sichuan white rabbits, Front. Genet. 12 (2021), 777232.
[78] B. Zhang, Z. Yan, P. Wang, et al., Identification and characterization of lncRNA and mRNA in testes of landrace and Hezuo boars, Animals 11 (8) (2021), 2263.
[79] GRA. Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences. accessed 2 September 2021;https://ngdc.cncb.ac.cn/gsa/.
[80] D. Bo, X. Jiang, G. Liu, et al., Multipathway synergy promotes testicular transition from growth to spermatogenesis in early-puberty goats, BMC Genomics 21 (1) (2020) 372.
[81] T. Li, R. Luo, X. Wang, et al., Unraveling stage-dependent expression patterns of circular RNAs and their related ceRNA modulation in ovine postnatal testis development, Front. Cell Dev. Biol. 9 (2021), 627439.
[82] W. Nolte, R. Weikard, R.M. Brunner, et al., Biological network approach for the identification of regulatory long non-coding RNAs associated with metabolic efficiency in cattle, Front. Genet. 10 (2019) 1130.

Biochim Biophys Acta Gene Regul Mech. 2022 May 2;1865(4):194815.

From the comparative study of a circRNA originating from an mammalian *ATXN2L* intron to understanding the genesis of intron lariat-derived circRNAs

Annie Robic [a,*], Chloé Cerutti [a], Julie Demars [a], Christa Kühn [b,c]

[a] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France
[b] Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), 18196 Dummerstorf, Germany
[c] Faculty of Agricultural and Environmental Sciences, University of Rostock, 18059 Rostock, Germany

DOI: 10.1016/j.bbagrm.2022.194815

A B S T R A C T

Circular intronic RNAs (ciRNAs) are still unexplored regarding mechanisms for their emergence. We considered the *ATXN2L* intron lariat-derived circular RNA (ciRNA-ATXN2L) as an opportunity to conduct a cross-species examination of ciRNA genesis. To this end, we investigated 207 datasets from 4 tissues and from 13 mammalian species. While in eight species, ciRNA-ATXN2L was never detected, in pigs and rabbits, ciRNA-ATXN2L was expressed in all tissues and sometimes at very high levels. Bovine tissues were an intermediate case and in macaques and cats, only ciRNA-ATXN2L traces were detected. The pattern of ciRNA-ATXN2L restricted to only five species is not related to a particular evolution of intronic sequences. To empower our analysis, we considered 221 additional introns including 80 introns where a lariat-derived ciRNA was previously described. The primary driver of micro-ciRNA genesis (< 155 nt as ciRNA-ATXN2L) appears to be the absence of a canonical "A" (i.e. a "tn**A**" located in the usual branching region) to build the lariat around this adenosine. The balance between available "non canonical-A" (no ciRNA genesis) and "non-A" (ciRNA genesis) for use as a branch point to build the lariat could modify the expression level of ciRNA-ATXN2L. In addition, the rare localization of the 2′-5′ bond in an open RNA secondary structure could also negatively affect the lifetime of ciRNAs (macaque ciRNA-ATXN2L). Our analyses suggest that ciRNA-ATXN2L is likely a functionless splice remnant. This study provides a better understanding of the ciRNAs origin, especially drivers for micro ciRNA genesis.

Table 2: Contingency table of introns and ciRNAs according to the branching region characteristics.

The availability of a "tn**A**" is statistically different within the group of 234 introns (and within the subgroups of 76 micro-introns, 71 mini-introns, 87 midi-introns) according to their involvement or not in ciRNA production (p-value < 0.001).

	76 micro-introns (90 < bp < 180)	71 mini-introns (181 < bp < 320)	87 midi-introns (330 < bp < 1,100)	234 introns considered
"tn**A**" available in the branching region	44/53	27/35	53/61	149 introns producing no ciRNA
	1/23	8/36	5/26	85 introns producing ciRNAs
120 ciRNAs considered and produced from 85 introns	37 micro-ciRNAs	51 ciRNAs	32 large-ciRNAs	
The nucleotide used as a branch point is not located in the usual branching region	0/37	0/51	1/32	
"A" is the nucleotide used as a branch point	0/37	1/51	2/32	