ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Graduate Programme in Molecular Biology and Biomedicine

Master Thesis

Exons going in circles: Computational approaches reveal universal genomic architecture design of the CircExons

MSc Student Candidate
Varamogianni Mamatsi Vassiliki
Supervision
Christoforos Nikolaou

October 16, 2020

Exons going in circles: Computational approaches reveal universal genomic architecture design of the CircExons

MSc Student Candidate

Varamogianni Mamatsi Vassiliki

Supervision

Christoforos Nikolaou

” [...] να μυριστεί κατά πού τραβάει η νιότη και να πλωρίσει κατά κει αυτός την ψυχή του.
Να πώς, ακολουθώντας τη διδασκαλία του μαθητή μου, έφτασα στην Κρήτη.”
-Ν. Καζαντζάκης

Contents

List of Figures

List of Tables

0.1 Acknowledgments

This master thesis was accomplished in the context of the master program ”Molecular Biology and Biomedicine”, which was the most important challenge ever had so far, while I got to know that all we know so far.. is never enough! Thus, I would like to thank all the Professors and the people who support this Master Program and make it possible every year.

I would also like to express my special thanks to:

My Professor Christoforos Nikolaou, who was willing to support and supervise my master thesis. I would like to thank him for the opportunity he gave me to work on human genomics and for the trust he showed me during the time we were collaborating.

Pavlos Pavlidis and Ioannis Iliopoulos for being my thesis co-supervisors.

All the people from Computational Genomics Group, for the great vibe and their help during all this time.

My beloved ”Charile’s Angels” Sofia Papanikolaou and Eleni Lianoudaki. We got through a very rough time together, but the team work, made the dream work.

Chrysa, Despina, Sofia and Thomi for their outstanding help on the present manuscript.

The ”Pigmies” and the ”Heraklion people” for making these years one of the most creative and lively periods of my life.

My parents Thanasis and Niki, as well as my sister Despina. For once again they believed in me more than I did and they supported me all this years in any way possible to get where I wanted to be.

0.2 Περίληψη/Abstract

Τα Κυκλικά εξόνια (CircExons) είναι μονόκλωνα μόρια RNA , κλεισμένα με ομοιοπολικό δεσμό και ανήκουν στην κατηγορία των κυκλικων ριβονουκλεικών οξέων (μη κωδικά RNAs . Ο μηχανισμός παραγωγής τους υποστηρίζεται πως είναι μια ειδική μορφή εναλλακτικού ματίσματος, που ονομάζεται Ανάποδο μάτισμα (Back-Splicing), όπου στην περίπτωση αυτή η μηχανή ματίσματος συρράπτει μια θέση δότη που βρίσκεται κάτωθεν της θέσης δέκτη με 3-5 φωσφοδιεστερικό δεσμό. Επιπλέον, τα κυκλικά εξόνια δεν παράγονται τυχαία, αλλά είναι σαφώς καθορισμένα και τα μετάγραφα των γονιδίων που παράγουν τέτοια μόρια είναι συγκεκριμένα και διαφέρουν μεταξύ κυτταρικών τύπων. Ωστόσο, οι παράγοντες που οδηγούν στην αυστηρή αναγνώριση και παραγωγή των κυκλικών εξονίων παραμένουν άγνωστοι. Στην παρούσα μελέτη, η οποία πραγματοποιήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος "Μοριακή Βιολογία και Βιοϊατρική" , παρουσιάζεται ένα κοινό γενομικό αρχιτεκτονικό πρότυπο που απαντάται σε διαφορετικούς κυτταρικούς τύπους του ανθρώπου και αφορά στην αλληλουχία του DNA και την τοποθέτηση των νουκλεοσωμάτων σε αυτό. Ειδικότερα, υποστηρίζουμε ότι τα κυκλικά εξόνια παρουσιάζουν χαμηλότερη περιεκτικότητα σε γουανίνη και κυτοσίνη (GC content), αλλά ισχυρότερη νουκλέοσωμική τοποθέτηση σε σύγκριση με τα γραμμικά εξόνια. Ταυτόχρονα προκύπτουν υποψίες ότι, τα άνωθεν των κυκλικών, γραμμικά εξόνια διαδραματίζουν καθοριστικό ρόλο στη σηματοδότηση των κυκλικών εξονίων που έπονται. Μέσα από την παρούσα διατριβή φαίνεται αυτά τα στοιχεία να αντιπροσωπεύουν έναν καθολικό γενομικό σχεδιασμό που διακρίνει τα κυκλικά εξόνια από τα γραμμικά εξόνια στα ανθρώπινα κύτταρα και ενδεχομένως να δικαιολογεί το φαινόμενο του ανάποδου ματίσματος έναντι άλλων μορφών εναλλακτικού ματίσματος.

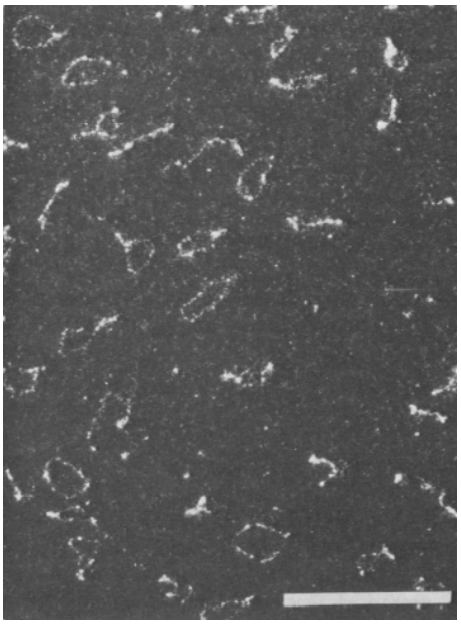
Circular Exons (or CircExons) are single stranded RNA molecules, covalenty closed and belonged to the category of circRNAs (non conding RNAs). The mechanism of their production is assumed to be a special form of RNA splicing, called Back-Splicing, where the splicing machinery ligates a downstream splice donor site reversely with an upstream splice acceptor site, forming a 3-5 phosphodiester bond. In addition, CircExons are not randomly produced

but each cell line produces specific CircExons. However, the mechanisms that result in this kind of strict recognition and production of the circular exons are remained unknown. In the present study, which was accomplished in the context of the Master Program ”Molecular Biology and Biomedicine”, is presented a genomic architecture pattern which corresponds to different human cell types and concerns DNA sequence and nucleosome positioning. In particular, here we support that CircExons present lower GC content but stronger nuclesome positioning than linear exons. Additionally, we believe that upstream linear exons (the circular ones) play a key role in signaling circular exons. These patterns seem to represent a universal genomic design for the human CircExon recognition and probably justifies the phenomenon of back splicing, instead of other forms of alternative splicing.

Chapter 1

Introduction

The first evidence for the existence of circular RNAs, was published in 1976 by Heinz L. Sänger and his team. They presented that potato spindle tuber viroid is a single-stranded, covalently closed, circular RNA molecule. Thus, the research concerning the comprehension of circRNAs was initiated (Sanger, Klotz, Riesner, Gross, & Kleinschmidt, 1976).



This picture was published along Sänger's observations, showing various types of RNA structures ranging from rods and dumb-bells to single-stranded circles (Electron Microscope,1975).

1.1 Linear splicing and back splicing

In eukaryotic organisms, the precussor mRNA (pre-mRNA), which includes introns and exons, will undergo splicing in order to exclude introns and form the mature mRNA. Usually, splicing is undergone in a linear manner, where exons are ligated and introns are removed (constitutive splicing). However, alternative splicing is the procedure which results in various transcripts of the same gene and thus, different isoforms of the same protein (Wang et al., 2015), due to different combinations of the consisted exons (or introns when intron retention occurs) of the gene (Figure 1.1. This different combination of exon stitching explains the size deviation in human and other eucaryotic organisms, between the proteome (protein counts) and the genome (the entire set of genes present in any cell or an organism). The main event of splicing is accomplished by the spliceosome, when the donor site is an upstream 5 splice site, the acceptor site is a downstream 3 splice site and those will join together in order to give rise in a linear mRNA.

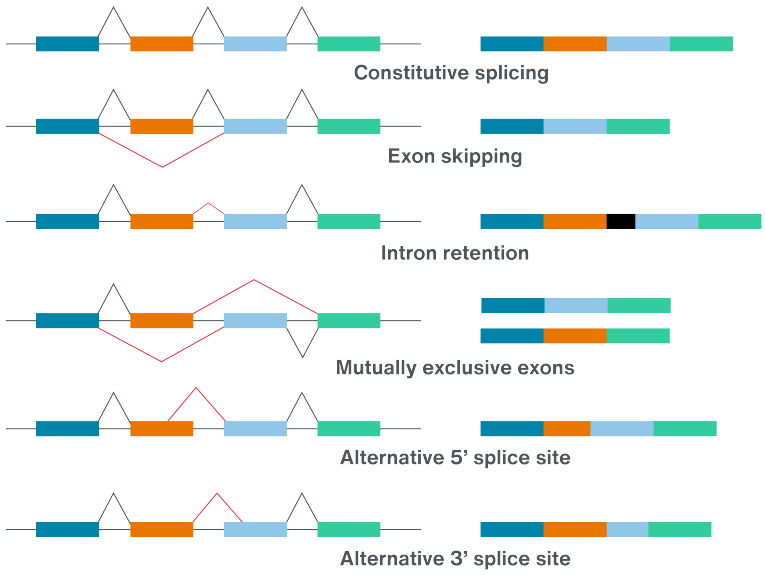


Figure 1.1: Alternative splicing produces different protein isoforms. Source: Nanoporetech.

Exons and introns are well defined and recognized by the spliceosome machinery, which scans the pre-mRNA with 5-end to 3-end orientation. The first characteristic of exons are consensus sequences of the donor and acceptor site of the junction. As it is described, an intron is marked by invariant sequences. In particular, the splice donor site includes an GU dinucleotide at the 5' end of the intron (intron start), while the splice acceptor site (intron end) is composed by a AG dinucleotide at the 3' end (Figure 1.2). Further upstream the 5'-end, there is a polypyrimidine tract and further upstream is located an Adenine, referred as the branchpoint, which is crucial for the splicing and the formation of the splice loop. (IIDA & SASAKI, 1983; Kitamura-Abe, Itoh, Washio, Tsutsumi, & Tomita, 2004; Burset, Seledtsov, & Solovyev, 2001).

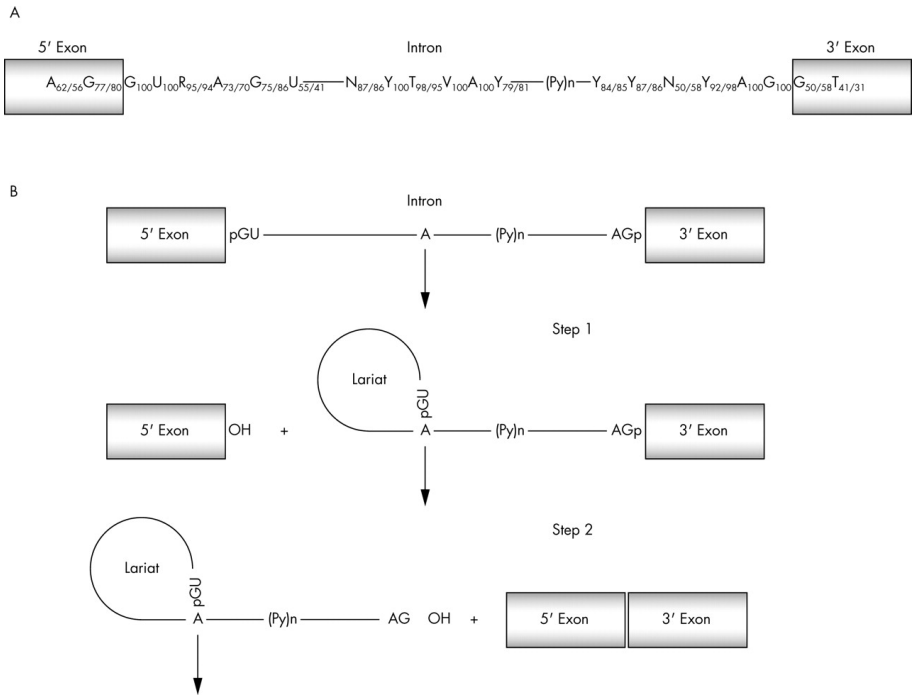


Figure 1.2: Pre-mRNA splicing

Unlike common alternative splicing and constitutive splicing, which occurs in a linear fashion and refers to mRNAs and lncRNAs, back splicing is a different type of splicing, where a downstream 5 splice (downstream exon end) site joins an upstream 3 splice (up-

stream **E**xon **S**tart **S**ite or ESS) by the spliceosome machinery and give rise to a circular product, with a 3,5' phosphodiester bond, at the back-splicing junction site (Zhang et al., 2016)(Figure 1.3). Even though the mechanism of the event is clear and both linear and back splicing require the canonical spliceosomal machinery (Starke et al., 2015), the reason why the spliceosome machinery chooses to back splice the exons remains unknown.

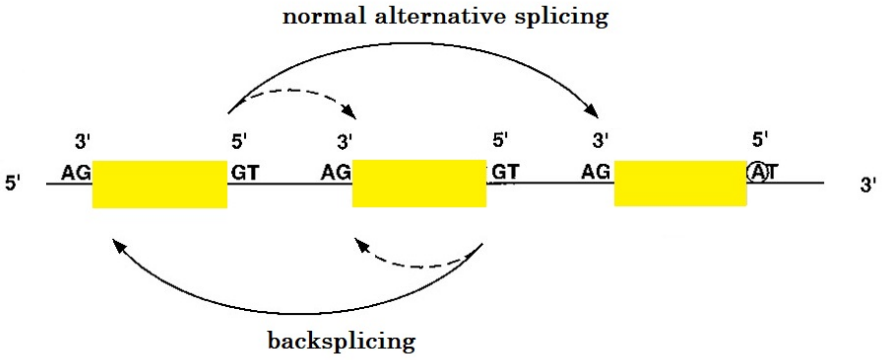


Figure 1.3: Back splicing mechanism VS Alternative splicing

1.2 Circular RNAs: A special category of non-coding RNAs

1.2.1 Characteristics of circRNAs

Most Circular RNAs or circRNAs are molecules originated from protein-coding genes with exonic or exonic-intronic composition, formed as covalently closed continuous loops and located in the cytoplasm of the cell (Jeck et al., 2013) (Figure 1.4). It was recently shown that circRNAs derived from adjacent exons, are more common than those derived from a single exon (Guo, Agarwal, Guo, & Bartel, 2014). circRNAs are typically not capped or polyadenylated, in contrast to their parental pre-mRNA, and are protected from exonucleases due to

covalent enclosure (Pamudurti et al., 2017).

Interestingly, both alternative splicing and back splicing and thus, linear and circular RNAs, are controlled by the spliceosome machinery (Starke et al., 2015). So far, Jeck et al. (Jeck et al., 2013) suggested two models of circRNA formation: 1) Circularization occurs through loop formation due to exon skipping and 2) circularization occurs through ALU complementarity or other RNA secondary structures between nonsequential donor-acceptor pairs. Either model could be vaild, however, further genomic exon definition rises unanswered questions in pre-mRNA splicing. Therefore, the identification of the characteristics of back-splice junctions are crucial for circRNA annotation and comprehension.

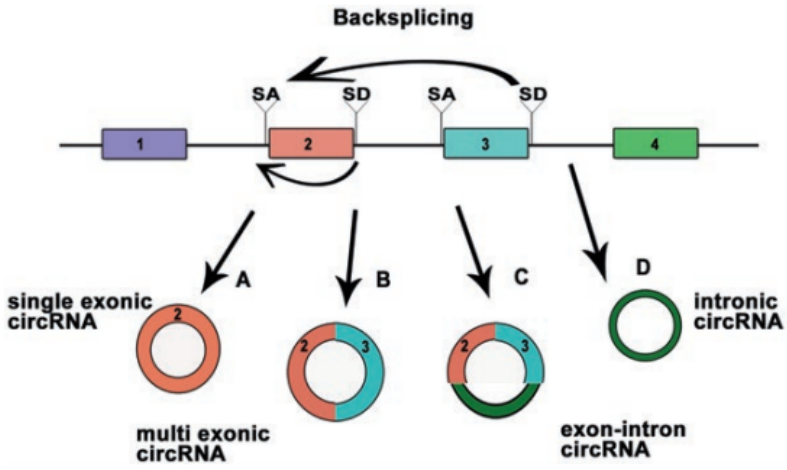


Figure 1.4: Types of Non Coding RNAs produced from Back Splicing

1.3 Function of circRNAs

1.3.1 circRNAs act as mi-RNA sponges

circRNAs are shown to be involved in two main functions in the cells. The first important paper released was in 1993 by Capel et.al, describing the finding of a single circular exon, for

Sry gene (which determines sex in mammals) in adult testis. This circRNA was present in the cytoplasm rather than the nucleus and untranslated (Capel et al., 1993). In 2013, Hansen and his co-workers explained that the circRNA of Sry, serves as a miR-138 sponge, with 16 target sites for miR-138. The function of microRNAs is to pair within mRNA sites, aiming to translational repression and mRNA destabilization (Bartel, 2009). They also showed that the circRNA ciRS-7 acts as a micro-RNA (miRNA-7) sponge, containing over 70 conserved miRNA target sites, and is related to Argonaute (AGO) proteins in a miR-7-dependent manner (Hansen, Jensen, et al., 2013). These circRNAs are also formed from exon back-splicing and are consisted exclusively of exonic sequences (Memczak et al., 2013)

1.3.2 circRNAs affect transcription machinery

Regulation of gene expression is depended basically on the affinity of RNA polymerase II with the DNA and the rest of the participating transcription factors. As mentioned above, a class of non-coding circular RNAs acts as miRNA sponges. However, another class which has been identified in HeLa cells (Z. Li et al., 2015), is correlated directly with the transcription machinery. These circRNAs are consisted of exons and middle intronic regions, and interact with the U1 subunit of the spliceosome machinery and thus, with the RNA polymerase II (Figure 1.5). In this work, pulldown assays were performed and it was shown that these circRNAs interact directly with the RNA polymerase II and with the U1 spliceosomal RNA of the U1snRNP. Along with these findings, it was proved that these circRNAs interact in cis with their parental transcripts. Interestingly, when the circRNAs were absent (by knocking down the relevant exons), the abundance of the parental mature mRNA was reduced. As described by Kwek, the subunit U1 interacts with the transcription factor TFIIH, which is vital for the initiation of the transcription, and collectively, induce the formation of the first 5' phosphodiester bond from the RNA polII (Kwek et al., 2002).

Transcription rate could be affected by nascent circRNAs processing. Even though back splicing is less favorable than canonical splicing due to low catalytic efficiency (the 5,3 phosphodiester bond of the back-splicing, is not sterically favored), back-splicing circularization

is associated with fast RNA polII elongation rate (Zhang et al., 2016). This is a fair conclusion because the spliceosome assembly is correlated with RNA polII and TFs (Martins et al., 2011).

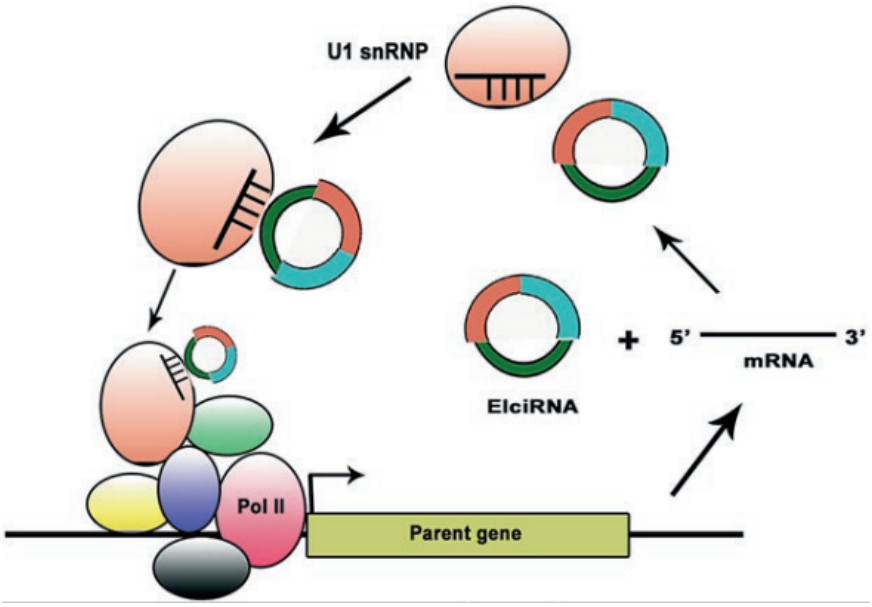


Figure 1.5: Regulation of transcription initiation by EIciRNA-U1 interaction. This complex interacts with the Pol II transcription initiation complex at the promoter of parent genes and promote their transcription (Kwek et al.,2002).

1.4 circRNAs in human cells

circRNAs are supposed to be crucial regulators in the human body and a series of works support that their expression is abundant (Rybak-Wolf et al., 2015; You et al., 2015; Hansen, Kjems, & Damgaard, 2013; Bachmayr-Heyda et al., 2015). In addition, circRNAs production is correlated with diseases like cancer (Hansen, Kjems, & Damgaard, 2013; Bachmayr-Heyda et al., 2015; Zhao, Cai, & Xu, 2019) and other disorders, as well as with development and differentiation. The role of circRNAs in human cancer includes their action as miRNA

sponges and also, as gene splicing, transcription, translation, or peptide and epigenetic regulators (Zhao et al., 2019).

Lately, research on circRNAs is focused on the production of circRNAs in human and mouse brains, especially in the cerebellum (Rybak-Wolf et al., 2015; You et al., 2015). As Rybak-Wolf and his team mentioned in their work, circRNAs are enriched in the cerebellum due to its great abundance in neural cells. Nevertheless, circRNAs are extremely stable molecules and thus, in neuronal cells that do not undergo mitosis, circRNAs are produced and tend to gather due to RNase resistance zhang2016biogenesis. Additionally, circRNAs are shown to be upregulated during neuronal differentiation and many of them act independently from their linear mRNAs, indicating that these molecules have a vital functional role in neurons (Rybak-Wolf et al., 2015; Gokoolparsadh, Anwar, & Voineagu, 2018). However, the mechanisms that result in their raised and independent production from the parental transcript, is not clear yet.

1.5 Nucleosomes

The haploid human genome is consisted of $3 * 10^9$ base pairs of DNA packaged into 23 chromosomes and thus, the diploid human genome contains $6*10^9$ base pairs of DNA per nucleus of somatic cell. This formidable folding, is achieved with the formation of nucleosomes in the nuclei of eukaryotic cells. The nucleosome is the fundamental subunit of DNA organization into chromosomes, containing in repeat two tetramers of histone proteins (H2A, H2B, H3, H4) and about 146 base pairs of wrapped DNA (Luger, Mäder, Richmond, Sargent, & Richmond, 1997) around histones creating a bead. Additionally, each of these beads is connected to each other with the histone protein H1 and linked DNA bases, resulting in a "beads on a string" motif of chromatin organization. Nucleosomes were first observed as particles in the electron microscope by Don and Ada Olins in 1974 (Figure 1.6).

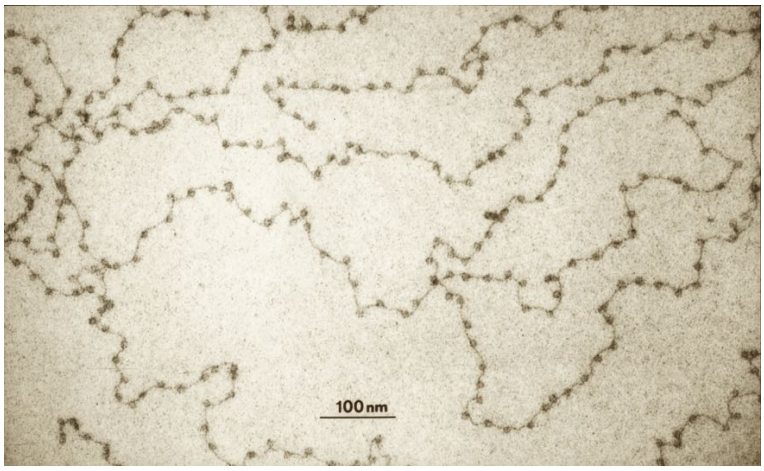


Figure 1.6: "Beads on a string". Image credit: Ada Olins and Donald Olins 1974, University of Tennessee/Oak Ridge Graduate School of Biomedical Sciences

1.5.1 Chromatin, Euchromatin and Heterochromatin

With the term Chromatin is described the DNA, RNA and protein complexes in the nucleus of the cell, forming the chromosomes. However, the DNA folding is not alike along its length. In contrast, chromatin could be densely packed, forming the heterochromatin complex, or loosely packed forming the euchromatin complex (Figure 1.7). The first scientist who observed differences along the chromosomes, was the botanist Emil Heitz in 1928 (Passarge, 1979) (Figure 1.8) who introduced the terms euchromatin and heterochromatin. Haitz stained chromosomes of different organisms and noticed two different stain intensities, corresponding to heterochromatin for strong intensity and euchromatin for milder.

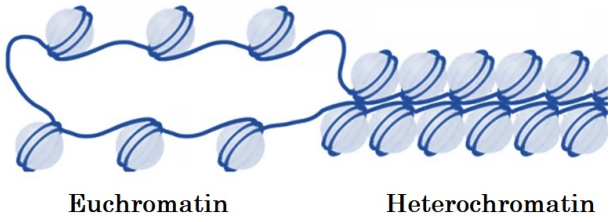


Figure 1.7: Euchromatin and Heterochromatin.

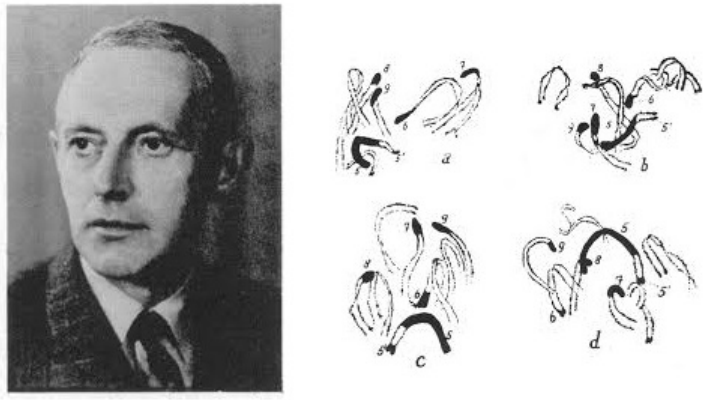


Figure 1.8: Left: Emil Heitz (1950). Right: Darkly stained heterochromatin and lightly stained euchromatin in *Pellia epiphylla* (Heitz, 1928). Source: Emil Heitz and the concept of heterochromatin, 1979.

Both heterochromatin and euchromatin are functionally and structurally distinct regions of the genome. In euchromatin regions, are located actively transcribed genes and the DNA is accessible to transcription factors. On the other hand, heterochromatin is less accessible to transcription factors and is trancriptionally inactive. For example, heterochromatic chromosomal regions are the telomeres and the centromere. Inactivation of the X chromosome in females occurs with heterochromatinization of the specific chromosome (Hennig, 1999).

1.5.2 Nucleosomes and gene expression

Nucleosomes are rather dynamically structured and are not only required for the folding and the creation of higher-order structures like chromosomes. In particular, studies support that chromatin structure must be modulated for gene expression, controlling the transcription, thus, the regulating gene expression (Wolffe, 1998; Luger, Dechassa, & Tremethick, 2012; Lai & Pugh, 2017; Chubb & Bickmore, 2003; Venkatesh & Workman, 2015; B. Li, Carey, & Workman, 2007).

Processes such as transcription require the two strands of DNA to come apart in order to allow polymerases and other transcription factors access to the DNA template (Alberts et al., 2018). Also, the nucleosomes included in actively transcribed genes are maintained, by transferring from the downstream to upstream DNA regions, upon RNA polymerase progression (Weintraub & Groudine, 1976). Consequently, changes in the nucleosome structure in order to make the DNA less or more accessible take place when RNA polymerase II transcribes through a nucleosome (Kujirai & Kurumizaka, 2020). An important study from Clarck and Felsenfeld in 1992, revealed that nucleosomes are maintained in the DNA, using a circular plasmid containing a unique nucleosome. The histone core was displaced and transferred to the other DNA regions on the circular plasmid template, after transcription by bacteriophage SP6 RNA polymerase (Clark & Felsenfeld, 1992).

1.5.3 Nucleosomes and splice sites: Exon start site

Revealing the detailed nucleosome organization along coding regions of DNA, crucial evidence for an even more comprehensive profiling of gene expression were provided. While the eukaryotic premature mRNAs undergo alternative splicing, Denisov and his partners in 1997 calculated expected nucleosome positions along the DNA strands, focusing on the regions around the splice junctions (exon-intron and intron-exon). Thus, the first evidence that the gene-splicing sites appear to be covered by nucleosomes was published (Denisov, Shpigelman, & Trifonov, 1997). Later on, in 2005, the sequence dinucleotide periodicity around splice junctions was published (YY and RR) (Kogan & Trifonov, 2005), suggesting that nucleosomes are linked with the sequences that direct intron removal.

In 2009, the research concerning the chromatin architecture and the splice sites intensified. Firstly, Schwartz described that the chromatin organization marks the exon–intron structure (Schwartz, Meshorer, & Ast, 2009), while Tilgner and Nikolaou studied the intensity of nucleosome occupancy related to exon splice site strength (a strong splice site is highly selected for splicing). Additionally, they included that stable nucleosome positioning is more frequent in exons than in the surrounding introns (Tilgner et al., 2009). Anders-

son et al. reported that nucleosomes are well positioned in exons and carry characteristic histone modifications (Andersson, Enroth, Rada-Iglesias, Wadelius, & Komorowski, 2009). All the studies above directly indicate the importance for chromatin organization in RNA processing.

1.6 GC content and nucleosome occupancy

Both Adenine (A) with Thymine (T) and Cytosine (C) with Guanine (G) undergo specific hydrogen bonding, in order to assemble the DNA core. A and T pair with a double hydrogen bond, while G and C with a triple hydrogen bond. At the level of the secondary nucleic acids structure, this extra hydrogen bond of the G-C pairing, offers additional stability, compared to the A-T. However, by taking into consideration large-scale folding of the DNA, which contains geometrical and steric constraints (tertiary structure), this triple hydrogen bond makes a GC-rich DNA chain more flexible and more stable, able to sharp bend.

The human genome (male) has 40.9% GC content (Piovesan et al., 2019), but gene-rich regions tend to have higher percentages than 40.9%. Since middle 1980, the first evidence for the correlation DNA sequence and nucleosome positioning (Satchwell, Drew, & Travers, 1986; Segal et al., 2006).

GC content has been characterized as an indirect recognition mechanism for DNA-protein binding sites. Firstly, Aeling et al. in 2007, published that regions with decreased DNA deformation energy (the energy required to bend DNA from its native shape to the conformation in a protein-bound complex) tend to wrap a protein (Aeling et al., 2007). In 2008 Kharchenko et al., supported that DNA deformation energy dips, indicate nucleosome binding hot-spots and are correlated with increased occurrence of GC-rich regions (Kharchenko, Woo, Tolstorukov, Kingston, & Park, 2008). Finally, Tilgner and Nikoalou found that the GC content pattern is similar to the pattern of nucleosome occupancy in human internal exons, by predicting the nucleosome occupancy with a tool called ”SymCurv” (Tilgner et al., 2009). The same year, Desiree Tillo and Timothy Hughes, using yeast MNase-seq data, also

published a direct correlation of the Nucleosome occupancy with the increased GC content in random yeast DNA compositions(Tillo & Hughes, 2009).

1.7 Aim of the study

The aim of the present study is to uncover universal architectonic patterns and motifs that characterise CircExons, compared with the linear exons resulting from normal splicing or other types of alternative splicing, concerning their sequence and their nucleosome occupancy. Additionally, this combination of analysis intends to reveal why these exons give rise to CircExons and whether the neighbour of these exons are affecting this determination. We have collected evidence which support a correlation between exon inclusion during alternative splicing and high nucleosome occupancy (Iannone et al., 2015) and that alternative splicing is influenced by chromatin architecture remodelling.

Moreover, exons and introns are distinct due to recognisable sequencing signals (recognised by the spliceosome and the transcription machinery). Tilgner and Nikolaou (Tilgner et al., 2009) correlated GC content and Nucleosome positioning in order to describe whether an exon is less or more favored for splicing. Thus, by analysing the GC content along with Nucleosome positioning of the CircExons, we concluded that CircExons differ from linear exons.

Chapter 2

Materials and Methods

2.1 Languages and overview

In this section will be analysed the methods that were applied in order extract the present results, concerning the genomic analysis of the circular Exons. For the needs of the particular study, the analysis performed mainly in R language and some applications run through Linux command line. Spearman correlation test also performed in R studio.

2.2 Data sets, cell lines and human reference genome

In the present work were analysed exons that potentially give rise to Circular Exons in human genome. In continuation, those exons compared with linear exons that do not give rise to circular elements. For this purpose, it was important to work with two different data sets:

• The first data set was in *.bed* format including the annotations of all human exons. This file was downloaded from the University of California Santa Cruz (UCSC) Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables), using the human reference genome hg19 (GRCh37) released in 2009/02/27 from the Genome Reference Consortium and tracking UCSC genes (*ALL_EXONS_hg19.bed*).

• The second data set was also in *.bed* format, including the CircExons' annotations expressed in different cell lines or tissues of the the human. Those files were downloaded from CIRCpedia_v2 (Dong, Ma, Li, & Yang, 2018) (https://www.picb.ac.cn/rnomics/circpedia/). The human genome hg19 was selected (Table 2.1) for the cell lines K562 (myelogenous leukemia cells), SH-SY5Y (human neuroblastoma cells), as well as Cerebellum cells (saved as *CE_[cell line]_hg19.bed*).

Cell lines	**Number of CircExons**
K562	27.403
SH-SY5Y	9.067
Cerebellum (multicell type)	27.473

Table 2.1: Cell lines and number of CircExons annotated

2.3 File conversion

If necessary, the *.bed* files were converted in *.fasta* (hg19 assembly) format, using the tool *bedtools getfasta*.

```
bedtools getfasta -fi hg19.fa -bed [exon type]_[cell line]_hg19.bed
```

2.4 Creation of Exons Categories and exon manipulation

2.4.1 Manipulation of the file containing all human exons

The file ALL_EXONS_hg19.bed includes information about all human transcripts and each annotation includes the exon coordinates as well as the exon count of the transcript and three helpful indexes were created (see below *txIndex, exonIndex, hg19Index*). The main goal was to unfold each transcript **(transcript unfolding)** in order to have a final file with exon annotations.

e.g: the first annotation of the table

```
  chrom txStart txEnd exStarts exEnds geneSymbol strand exCount
chr1 11873 14409 11873,12612,13220, 12227,12721,14409, DDX11L + 3
```

• **txIndex** (Transcript index)
Before the "transcript unfolding", each transcript was indexed. The *.bed* file was sorted according to the chromosome and the coordinates of each annotation using the following command:

```
sort ALL_EXONS_hg19.bed -k1,1 -k2,2n
```

• **exIndex** (Exon index)
Using the exon count (the number of exons in every transcript) of each annotation, an exon index was created, indicating the exon rank in the transcript (first exon, second, third etc).

e.g: the first three annotations of the table before exon indexing before transcript unfolding

chrom	txStart	txEnd	exStarts	exEnds	geneSymbol	strand	**exCount**	**txIndex**
chr1	11873	14409	11873,12645,13220,	12227,12697,14409,	DDX11L1	+	**3**	**1**
chr1	11873	14409	11873,12645,13220,	12227,12697,14409,	DDX11L1	+	**3**	**2**
chr1	11873	14409	11873,12594,13402,	12227,12721,14409,	DDX11L1	+	**3**	**3**

e.g: the exon annotations after transcript unfolding with the indexes

chrom	exStarts	exEnds	geneSymbol	strand	**exCount**	**txIndex**	**exIndex**
chr1	11873	12227	DDX11L1	+	**3**	**1**	**1**
chr1	12612	12721	DDX11L1	+	**3**	**1**	**2**
chr1	13220	14409	DDX11L1	+	**3**	**1**	**3**
chr1	11873	12227	DDX11L1	+	**3**	**2**	**1**
chr1	12645	12697	DDX11L1	+	**3**	**2**	**2**
chr1	13220	14409	DDX11L1	+	**3**	**2**	**3**
chr1	11873	12227	DDX11L1	+	**3**	**3**	**1**
chr1	12594	12721	DDX11L1	+	**3**	**3**	**2**
chr1	13402	14409	DDX11L1	+	**3**	**3**	**3**

• **hg19Index** The final index that the exons got, was according to the exon rank in the file (after the transcript unfolding) from 1 to the total exon count.

final format of the file ALL_EXONS_hg19.bed

chrom	exStarts	exEnds	geneSymbol	strand	**exCount**	**txIndex**	**exIndex**	**hg19Index**
chr1	11873	12227	DDX11L1	+	**3**	**1**	**1**	**1**
chr1	12612	12721	DDX11L1	+	**3**	**1**	**2**	**2**
chr1	13220	14409	DDX11L1	+	**3**	**1**	**3**	**3**
chr1	11873	12227	DDX11L1	+	**3**	**2**	**1**	**4**
chr1	12645	12697	DDX11L1	+	**3**	**2**	**2**	**5**
chr1	13220	14409	DDX11L1	+	**3**	**2**	**3**	**6**
chr1	11873	12227	DDX11L1	+	**3**	**3**	**1**	**7**
chr1	12594	12721	DDX11L1	+	**3**	**3**	**2**	**8**
chr1	13402	14409	DDX11L1	+	**3**	**3**	**3**	**9**

It should be mentioned that after the transcript unfolding, some exons appear multiple times (while are part of multiple transcripts). However it was very important to maintain the rank after the transcript unfolding in order to identify the exact origin (parental transcript) of the CircExon.

2.4.2 Pipeline for the creation of exon categories

The categories presented in Table 2.2 and Table 2.3 were created by manipulating the file CE_[cell line]_hg19.bed, which includes the CIRCpedia_v2 annotations and the annotations of the file ALL_EXONS_hg19.bed with the tool *bedtools intersect*.

Exon Type	**Description**
Circular Exons (**CE**)	CIRCpedia v2 annotations
Linear Exons (**LE**)	Exons excuded from CIRCpedia v2 annotations

Table 2.2: Exon Categories

Exon Type	**Description**
CE	Circular Exons as as described in Circpedia v2
CEi	Exons contained in Circular RNAs
LEi	Linear Exons contained in transcripts giving rise to circular RNAs
LEo	Linear Exons contained in transcripts which do not give rise to circular RNAs
LE	All linear exons (control)
AE	All exons (control)

Table 2.3: Exon Categories

```
   #Get CEi --exons included in the CircExon
bedtools intersect -a CE_[cell line]_hg19.bed -b ALL_EXONS_hg19.bed

   #Get LE --linear exons
bedtools intersect -a CE_[cell line]_hg19.bed -b ALL_EXONS_hg19.bed -v
```

Each annotation track of the output contained additionally the field *Exon Type* (CircExon for the CEi and LinearExon for the LE).

The files CEi_[cell line]_hg19.bed and LE_[cell line]_hg19.bed were merged and shorted according to the hg19Index. The new version of ALL_EXONS_hg19.bed (named ALL_EXONS_[cell line]_hg19.bed) describes whether an exon annotation is circular or linear.

In order to identify the Linear Exons that belong to transcripts that give rise to Circular Exons (LEi), it was necessary to get the txIndex from the file CEi_[cell line]_hg19.bed and find the annotations from ALL_EXONS_hg19.bed that match with this index. The annotations characterized as *LinearExon* correspond to the category LEi.

Finally, the Linear Exons that belong to transcripts that do not give rise to Circular Exons (LEo), were extracted with the tool *bedtools intersect*.

```
#Get LEo
bedtools intersect -a LE_[cell line]_hg19.bed -b LEi_[cell line]_hg19.bed
-v
```

The exon categories displayed in Figure 3.7 were created, by using an algorithm, which was built in order to register:
1) The downstream linear exons of each CircExon as N+1 and N+2.
2) The upstream linear exons of each CircExon as -2 and -1.
3) The first and the last exon of the CircExon as +1 and N in respect.

Special parameters:
1) In Single-Exonic CircExons the unique exon was annotated twice, as +1 (first exon) and as N (last exon):

```
chr1   980738   980903   AGRN   +   14   36   799   LinearExon   112   -2
chr1   981112   981256   AGRN   +   15   36   800   LinearExon   112   -1
chr1   981343   981468   AGRN   +   16   36   801   CircExon   112   +1
chr1   981343   981468   AGRN   +   16   36   801   CircExon   112   N
chr1   981539   981645   AGRN   +   17   36   802   LinearExon   112   N+1
chr1   981776   982115   AGRN   +   18   36   803   LinearExon   112   N+2
chr1   982199   982337   AGRN   +   19   36   804   LinearExon   112   L
```

2) Transcripts giving rise to more than one CircExon could contain exon annotations with dual characterizations, according to the relative position to each CircExon:

```
chr1  21132784  21133993  EIF4G3  -  26  26  9964  LinearExon  1079  N+1
chr1  21137230  21137377  EIF4G3  -  25  26  9965  CircExon    1079  N
chr1  21139650  21139732  EIF4G3  -  24  26  9966  CircExon    1079  +1
chr1  21143884  21144031  EIF4G3  -  23  26  9967  LinearExon  1079  N+1
chr1  21143884  21144031  EIF4G3  -  23  26  9967  LinearExon  1079  -1
chr1  21151592  21151691  EIF4G3  -  22  26  9968  CircExon    1079  N
chr1  21154109  21154191  EIF4G3  -  21  26  9969  CircExon    1079  C
```

All exon categories were analysed around their ESSs and **E**xon **E**nd **S**ites (EES) in Section 3.4, taking under consideration the strand of their parental transcript. The distances upstream and downstream the start sites as well as from their EESs were equally extended +/-500bp (the final length of each exon was 1000bp).

2.5 Nucleosome positioning analysis

The Nucleosome distribution analysis performed using two different approaches. A real time nucleosome positioning (using MNase-seq datasets) and a prediction model (using SymCurv (Tilgner et al., 2009)). The MNase-seq dataset (Experiment: ENCSR000CXQ) for the cell line K562 (*Homo sapiens*) was downloaded from ENCODE project (https://www.encodeproject.org/). The predictions for hg19 were extracted using SymCurv and the analysis performed in *.bed* files for each exon category. Finally, the scores for each exon were summarised in 100 bins (each bin contained 10bp of the total 1000bp around each ESS (or EES in respect) and plotted. Both MNase-seq data and SymCurv predictions, were used in a bigWig format. All SymCurv scores extracted were transformed:

$$\_SYMCURVscore_{final} = SYMCURVscore * 10$$

2.6 GC content analysis

The GC content around ESS (or EES as referred in Section 3.4) for the categories displayed in Table 2.3 and Table 2.2 was extracted from *.fasta* files.

2.7 Venn diagram

For the creation of the Venn diagram (Figure 3.20), the annotations of CIRCpedia_v2. for each cell type (K562 and Cerebellum) were intersected using *bedtools intersct* with file ALL_EXONS_[cell line]_hg19.bed of the other. In the command line was applied *-wa* in order to get the original entry of the -a.

```
bedtools intersect -a ALL_EXONS_[Cerebellum]_hg19.bed -b CE_[K562]_hg19.bed
-wa
bedtools intersect -a ALL_EXONS_[K562]_hg19.bed -b CE_[Cerebellum]_hg19.bed
-wa
```

Thus, it was possible to get the information for all CEi of each cell line, containing the characterization of Linear or Circular Exon. If an annotation from the output of the following command:

```
bedtools intersect -a ALL_EXONS_[Cerebellum]_hg19.bed -b CE_[K562]_hg19.bed
-wa
```

is characterized as LinearExon, it means that in K562 it is a circular exon but in Cerebellum this exon is linear.

2.8 Introns and Transcripts length

The length of introns and transcripts was calculated and plotted after log2 transformation.

Chapter 3

Results

This study aims to the genomic characterization of the formation of the circRNAs (in the present thesis will refer as **CircExons**, while exonic circRNAs were studied, containg exonic and intronic sequences). The term CircExon, refers to the product of every single back splicing event and thus, a CircExon could be Multi-Exonic (ME) or Single-Exonic (SE). Additionally, some transcripts of human genome do not give rise to CircExons and different cell type produces different CircExons, due to tissue specificity and differential gene expression.

In this section are displayed a series of bioinformatic analysis, which outline a genomic determination of the formation of CircExons, including Nucleosomic predictions and GC content analysis, comparing exons that potentially could be included in a back splicing event as well as exons undergoing normal alternative splicing in K562 cell line, human neuroblastoma SH-SY5Y cells as well as in Cerebellum. The data for the CircExons are collected by the online atlas Circpedia v2 (Dong et al., 2018).

3.1 Nucleosome occupancy is stronger in Linear Exons than in Circular Exons when predicted with *SymCurv*

Exon Start Sites (ESS) are important for splicing (Tilgner et al., 2009; Iannone et al., 2015). Thus, the first thing investigated, was whether CircExons (CE) and Linear Exons (LE) are distinct from each other according to the nucleosomic patterns around their start sites. It was hypothesised, that back splicing junctions differ from those of the normal alternative splicing, while back splicing and normal alternative splicing have opposite donor and acceptor sites. For this purpose, exons taking part in CircExons (expressed in K562 cells) were compared to Linear Exons, using the tool SymCurv. (Tilgner et al., 2009) This model predicts nucleosomic occupancy according to the structural property of natural nucleosome forming sequences. It was shown that nucleosome occupancy is stronger in linear than in circular exons (Figure 3.1, Table 3.1). However, Circular Exons present edgier nucleosome peak, while the Linear Exons score starts to increase more than 300bps upstream the ESSs.

In order to investigate the behaviour of Linear and Circular exons in more detail, human exons were classified in 5 categories (see **Materials and Methods** , Table 2.3).

All Linear Exons (LE) were divided in two subcategories, according to whether the original transcript gives rise to CircExons (LEi) or not (LEo) (Figure 3.2). Additionally, exons taking part in CircExons were examined independently (CEi) from the CIRCpedia_V2 annotations (CE). All exons (AE) of human genome and LE were used as control groups. The theoretical prediction model SymCurv, revealed that Linear Exons have in general stronger nucleosome occupancy whether they come from transcripts giving rise to CircExons or not (Figure 3.3). Again, all categories of Linear exons note an early increment of SymCurv Score and a mild dip before the ESS. In contrast, CE and CEi have an early decrement accompanied by a more sharp dip.

In continuation, a MNase-seq dataset was downloaded via ENCODE (https://www.encodeproject.org/) for the cell line K562 in order to get real time nucleosome signal (Figure 3.4). MNase-seq identifies nucleosome-dense regions. Interestingly, the results from MNase-seq analysis were different from those reported by SymCurv. CircExons (CE and CEi) that have more stable nucleosomes than Linear exons (LEi, LEo, LE) while their peak is notably higher. All exons (AE) of human genome were plotted

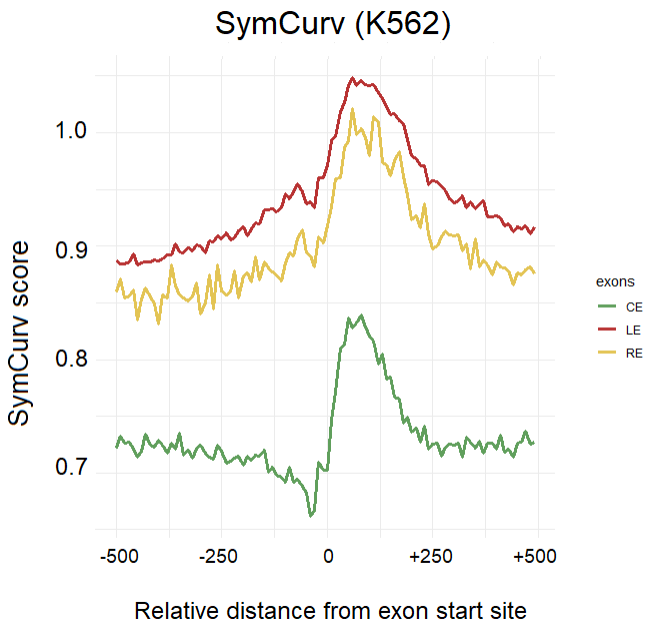


Figure 3.1: Nucleosome positioning prediction using SymCurv. **CE**: Circular Exons, **LE**: Linear Exons, **RE**: Random Exons.

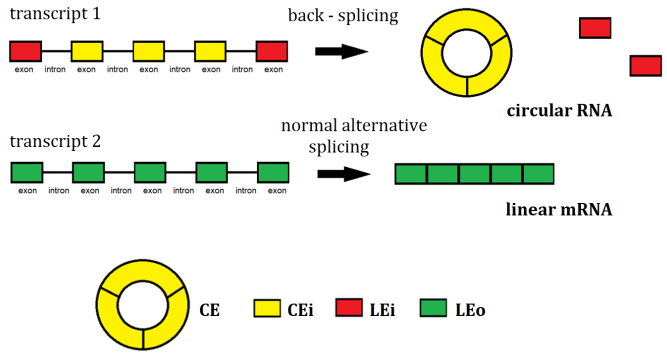


Figure 3.2: Schematic representation of exons types (Table 2.3).

as a control group and their pattern resulted more similar with this of Linear Exons. This was expected while linear exons are significant more abundant than circular. 'However, these opposite results raised

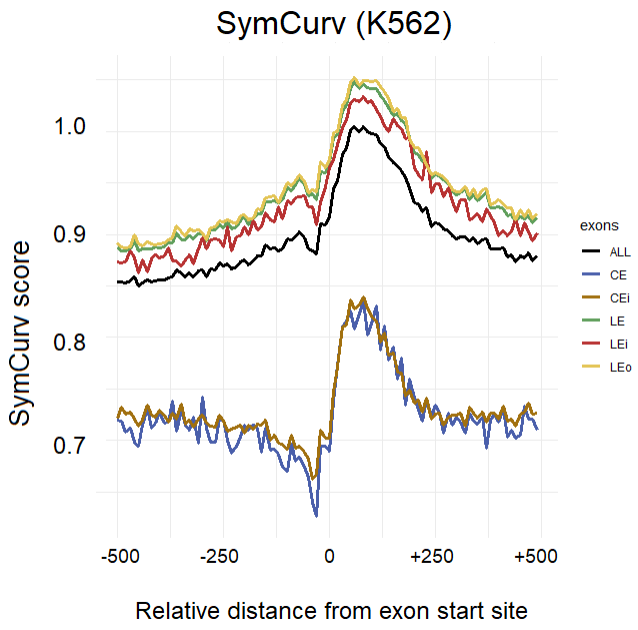


Figure 3.3: Nucleosome positioning prediction using SymCurv.

new critical questions as: *"Why the theoretical model supports that Linear Exons tend to be consisted by sequences more "nucleosome friendly" with higher GC content, while the MNase-seq signal is stronger in CircExons?"*

Exon Type	**min Value**	**max Value**
Circular Exons	0.6619576	0.8392911
All Linear Exons	0.8830589	1.0485
Random Exons	0.8323781	0.9986854

Table 3.1: SymCurv values for Circular Exons and Linear Exons.

Previous findings support that exon skipping is a phenomenon which leads to the formation of CircExons (Kelly, Greenman, Cook, & Papantonis, 2015) and also that exon skipping is correlated with low nucleosome densities (Kelly et al., 2015; Iannone et al., 2015).

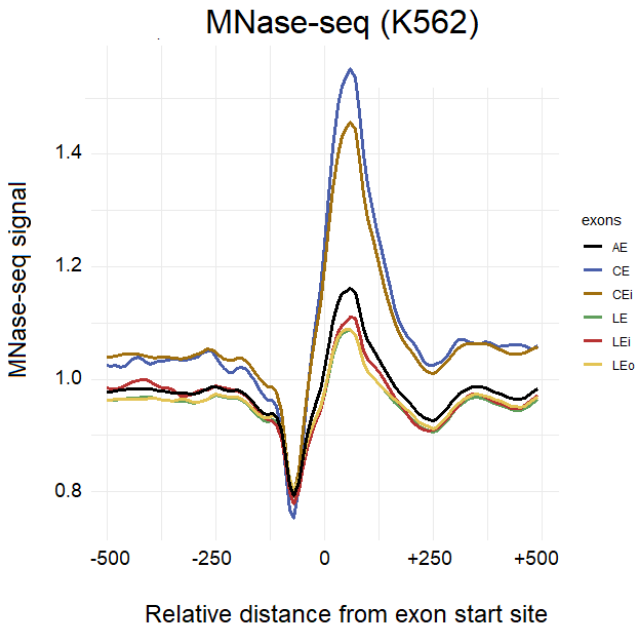


Figure 3.4: Quantitative MNase-seq maps nucleosome occupancy levels.

Exon Type	**min Value**	**max Value**
CE	0.6268659	0.8365137
CEi	0.6619576	0.8392911
LEi	0.8830589	1.034186
LEo	0.8629618	1.0485
LE	0.8867008	1.052049
AE	0.849922	1.004479

Table 3.2: SymCurv values for all exon categories.

3.2 GC content fluctuation distinguishes Linear and Circular exons.

It is clear that both in CircExons (CE) and exons taking part in CircExons (CEi), the GC content start to drop approximately 350bp upstream their start site, representing gradual fall of the guanine and citosine percentage (Figure 3.5). The lowest value of GC content (0.32) can be noticed upstream the ESS and few base pairs upstream the ESS, the content reaches the highest value (0.46). On the other hand, linear exons have stable GC content upstream the start site followed by very sharp drop of the content. In respect, the highest value for the linear exons is 0.53 (Table 3.3).

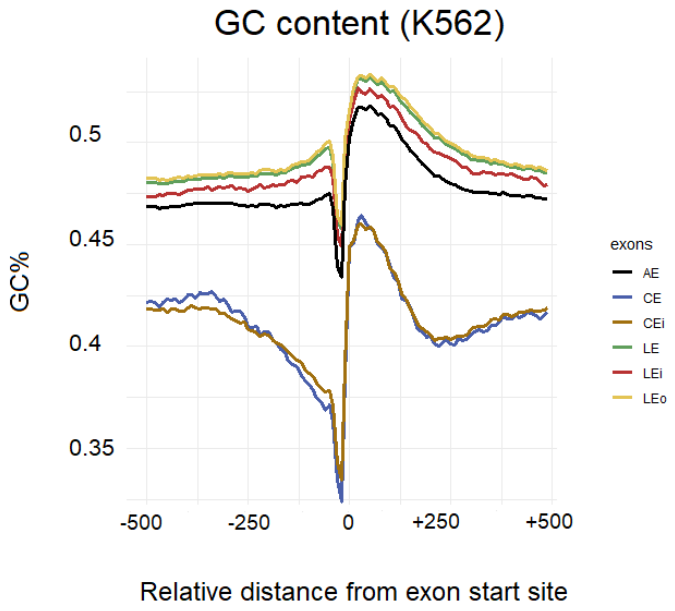


Figure 3.5: GC content around exon start sites.

This pattern seems to be in agreement with the pattern resulted from the SymCurv (Figure 3.3. Linear Exons have overall higher GC content and SymCurv score than Circular Exons, while the peak has a larger amount of tailing in Circular Exons than in Linear Exons. However, the transition to the GC content dip as well as the SymCurv score dip starts earlier in Circular Exons. Thus, upstream each CircExon start site (which is determined by sharp increasement of the GC content) the GC content start to fall, while in Linear exons the GC content remains stable.

Exon Type	**max Value**	**min Value**
CE	0.4641586	0.3237379
CEi	0.4617143	0.3320890
LEi	0.5312539	0.4504407
LEo	0.5251385	0.4400895
LE	0.5334471	0.4541526
AE	0.5145984	0.4228132

Table 3.3: GC percentage values for all exon categories.

3.3 Linear exons upstream the CircExons are probably crucial for CircExon recognition.

By performing statistical analysis, it was observed that the majority of the first exons present in CircExons, are the second exons of their parental transcript. Also, there is a linear negative correlation (-0.95) between the exon index (EI) of the first exon and the preference frequency, as Spearman correlation test indicated (in general, primal exons are favoured, than terminal).

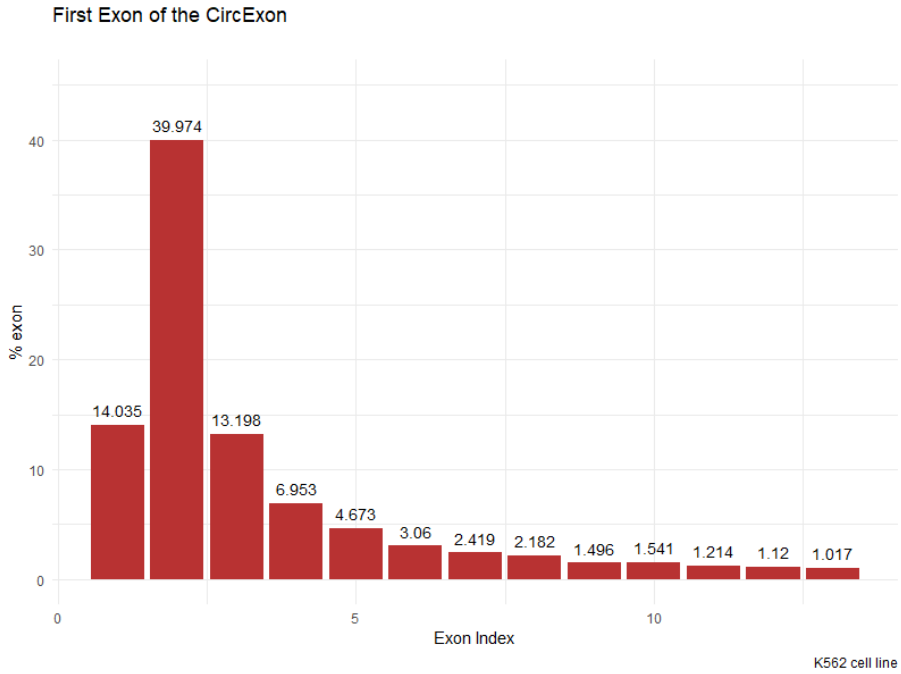


Figure 3.6: Exon Index of the First Exon included in CircExons.

Thus, once the differences between circular and linear exons were identified, the following step was to investigate the neighborhood of the first exon of the CircExon (Figure3.7).

It was observed, as expected, that +1 exon, which corresponds to a circular exon, had lower Symcurv score (Figure 3.8) than the -1 exon (linear exons) which had higher score and thus stronger nucleosome occupancy, indicating an abrupt change that might be important for the CircExon recognition by the transcription machinery. Again, MNase-seq analysis showed that +1 exon had higher

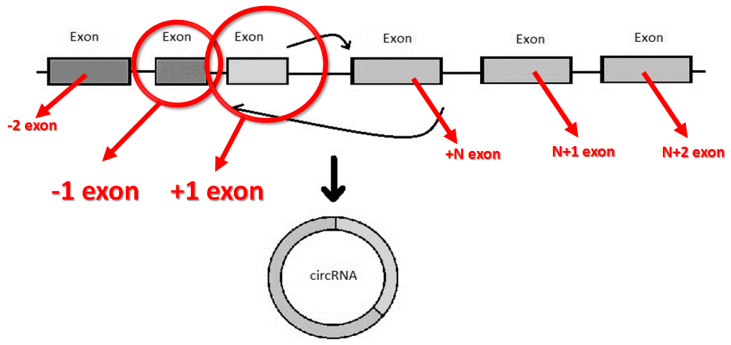


Figure 3.7: Schematic representation of the CircExon's neighbourhood.

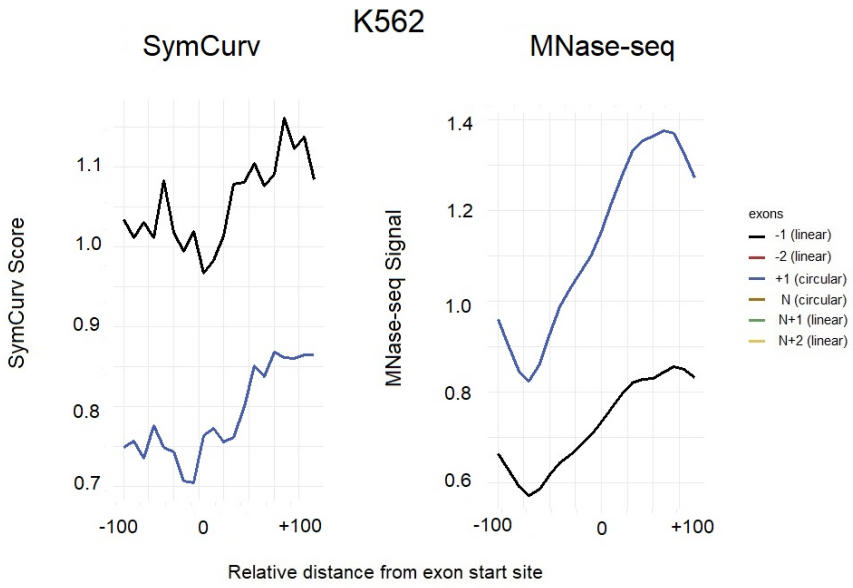


Figure 3.8: Comparisson between +1 and -1 exons. Right: MNase seq analysis. Left: SymCurv analysis.

scores than -1 (Figure 3.9).

Hence, as described in Figure 3.9, more exons of the neighbour were analysed, in order to identify the fluctuation of the nucleosome occupancy among exons of the CircExon (+1, N), as well as of the upstream and downstream linear exons (-2, -1, N+1, N+2) (Figure 3.7. Indeed, while the

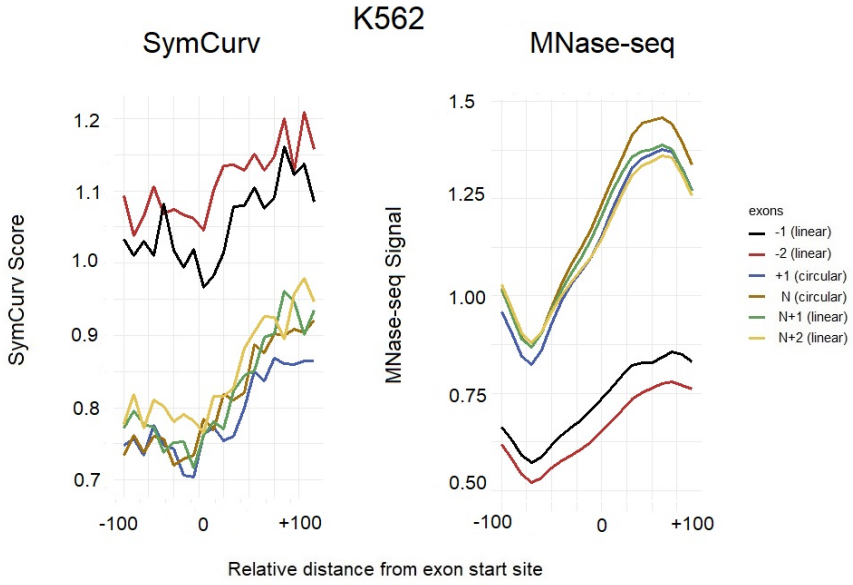


Figure 3.9: SymCurv and MNase-seq analysis for the exons consisting the neighborhood of the CircExon.

upstream linear exons seem to have different nucleosomic characterization from the circular exons, the downstream linear exons (N+1, N+2) own the motifs of the CircExons.

Additionally, the GC content (Figure 3.10) and the motifs were in agreement to the nucleosomic patterns of the SymCurv. Upstream linear exons (-1 and -2) seem to have more stable GC percentage which is related with robust nucleosomes while the CircExons and the downstream linear exons have wider fluctuations in GC content, which is related with loose nucleosome positioning. Another very important observation, was that the GC content actually, follows a pattern along the transcript. In particular, the GC content is increasing when approaching the first exon of the CircExon (GC%(-1) >GC%(-2)), which owns the lowest values. Downstream the circle, the content gradually starts to rise again.

So here, a key role of the GC content is presented which is represented by characteristic GC content displacement of the downstream linear exons of the CircExon and significant higher and more stable GC content of the upstream linear exons.

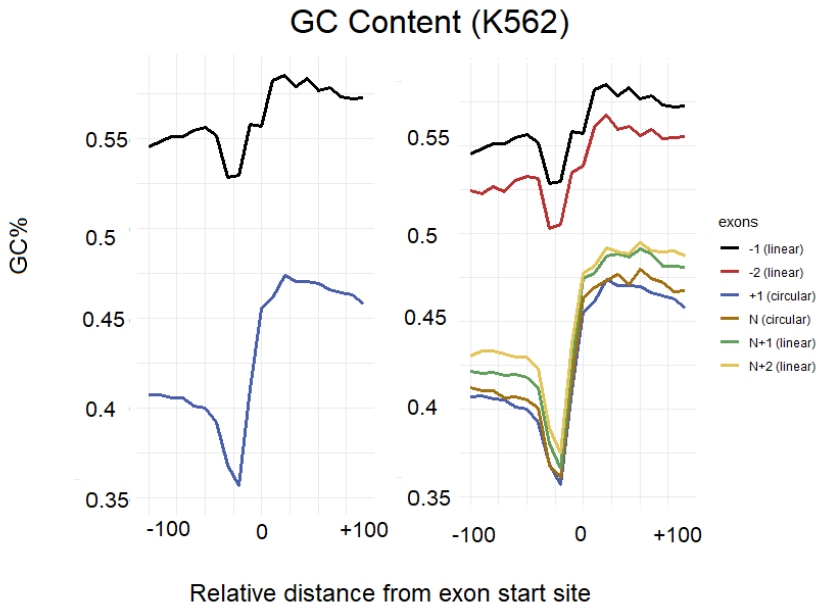


Figure 3.10: GC content around ESS.

Exon Type	**# exons**	**max Value**	**min Value**
-2 (linear)	7616	0.57	0.50
-1 (linear)	15409	0.58	0.53
+1 (circular)	17501	0.47	0.36
+N (circular)	16493	0.48	0.36
N+1 (linear)	13989	0.49	0.37
N+2 (linear)	8737	0.49	0.38

Table 3.4: GC content minimum and maximum values for each exon categories.

3.4 The upstream and downstream exons of the CircExons have distinct characteristic around their EES.

There was a comparison of the GC content between exons that are included in the CircExon (referred in this study as CEi) and all linear exons (LE). Interestingly, the GC content values of linear exons are overall grater than the values of circular exons. On the other hand, as mentioned in Section 3.2 where ESSs were analysed, the peak of the circular exons have a larger amount of tailing than linear exons accompanied by gradual fall (before the tailing) and gradual increament (after the EES). Along this observation, right on the EES of the linear exons, there is a GC content short but edgy dip, which is not present in circular EES (Figure 3.11).

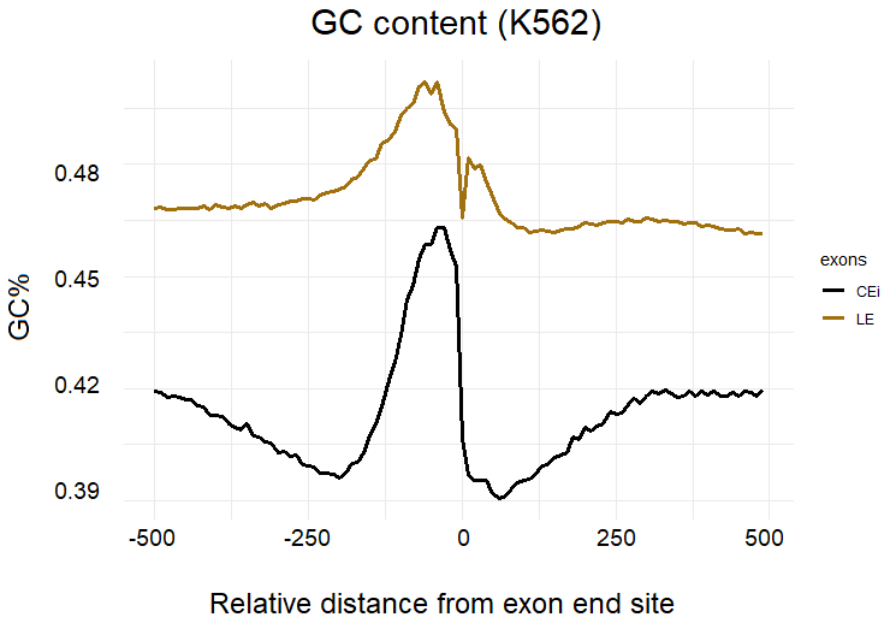


Figure 3.11: GC content of linear exons (LE) and circular exons (CEi) around EES.

In continuation, the intention was to define distinct GC content patterns between the upsteram and downstream linear exons that bracket the CircExon (Figure 3.7). In the above figure (3.12) it is shown that, the values range between upstream and downstream linear exons are distinct as mentioned in previous sections. The upstream exons are represented by greater values than the downstream, which are accompanied by the CircExons' values.

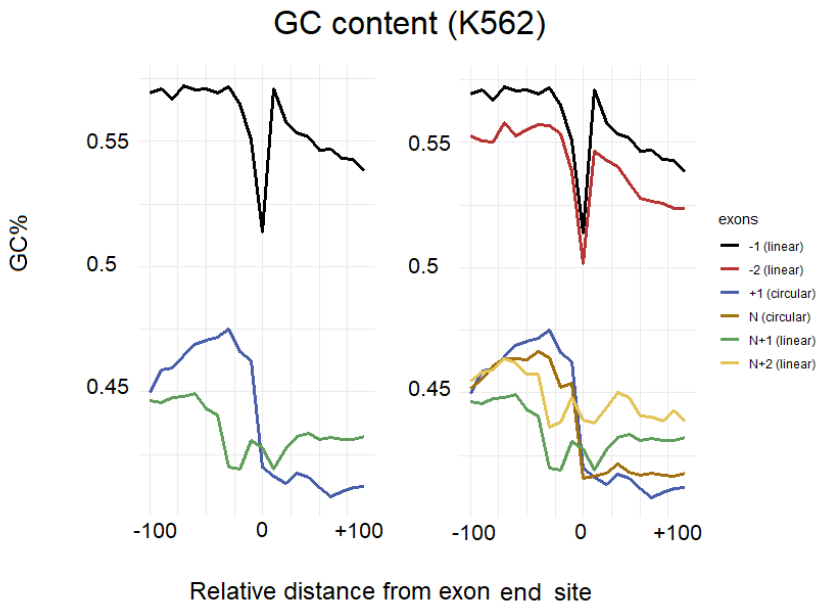


Figure 3.12: GC content analysis between EES of upstream linear exons (-1,-2), circular exons (+1,N) and downstream linear exons (N+1,N+2).

However, downstream linear exons (N+1, N+2) and circular exons here, seem to have different GC patterns. In particular, GC content of downstream exons has milder fluctuations before (N+2 pattern is slightly displaced from the N+1 pattern but seem to have similar motif) and after EES, while the GC content of the circular exons drops right on the EES as in Figure 3.11. Finally, the upstream exons have the sharp GC content dip on the EES which characterized previously the linear exons.

Finally, MNase-seq analysis (Figure 3.13) revealed that each of the three categories (upstream linear exons, circular exons and the downstream linear exons) own unique patterns.

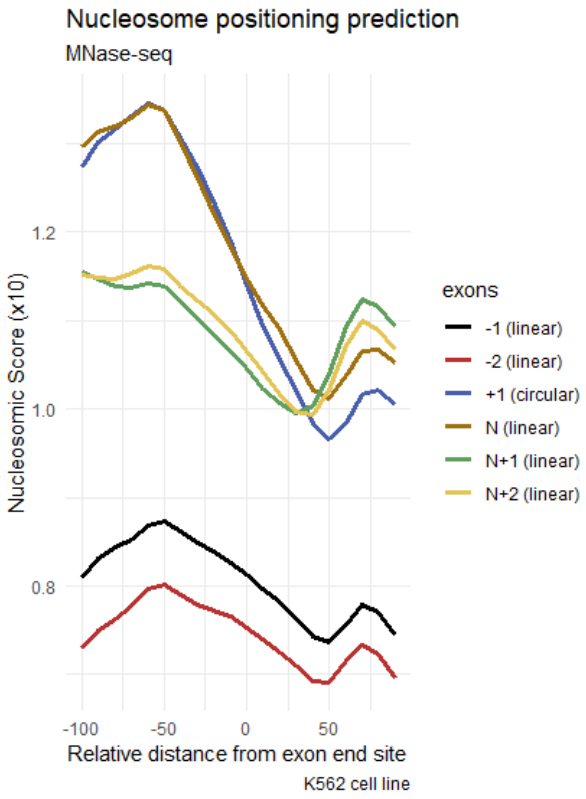


Figure 3.13: MNase seq analysis between exons of the neighbourhood.

3.5 Transcripts giving rise to CircExons tend to be longer than the average transcripts.

There was a significant difference when comparing the lengths of the transcripts that give rise to CircExons with those that do not. The first (Table 3.5) are significant longer while the mean value is 70825.74bp (median=41598, sd= 93995.02). The mean value for transcripts unrelated to CircExons is 44521.98bp (median=11523, sd=112539.1) and the average transcript length is 52428.64bp (median=18497, sd=107977.4). Those values are plotted in Figure 3.14. These findings confirm the published data by Kelly et al.,2015 who also supported that transcripts related to CircExon formation are longer.

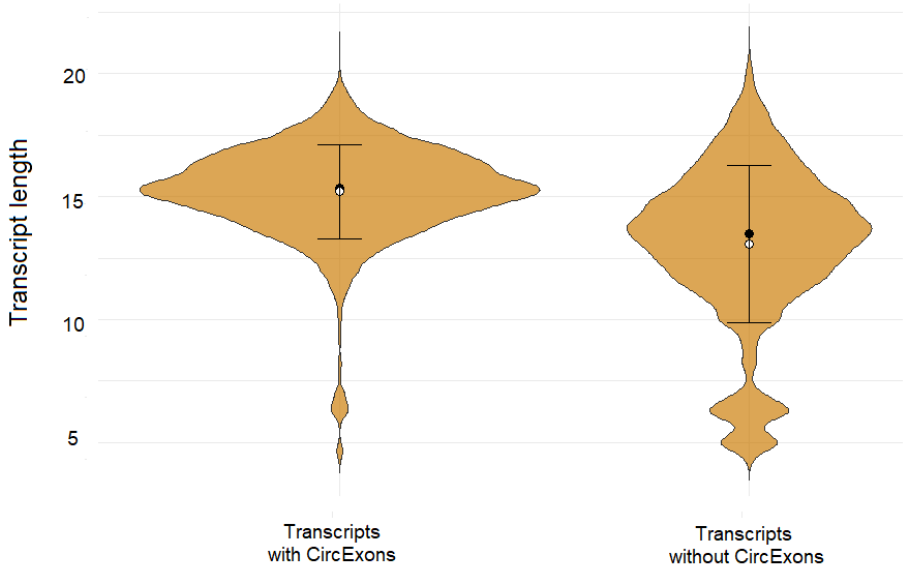


Figure 3.14: Transcripts giving rise to CircExons tend to be longer than the average transcripts. All values are log2-transformed.

Transcripts	Mean	Median	Standard deviation
with CircExons	70825.74	41598	93995.02
without CircExons	44521.98	11523	112539.1
average	52428.64	18497	107977.4

Table 3.5: Statistics of transcripts.

3.6 Introns bracketing CircExons are longer than internal introns.

Ashwal-Fluss and his co-workers mentioned that long introns are believed to be less favoured for splicing (Ashwal-Fluss et al., 2014). Additionally, in 2008, Roy and Kim enriched the already excisted theory (Fox-Walsh et al., 2005), that exon skipping which is closely related with circularization, is probably depended on the existence of long introns (Roy, Kim, Xing, & Lee, 2008).

Figure 3.15: Introns bracketing circExons are longer than introns included in the circExon. All values are log2-transformed.

Thus, our hypothesis was that introns bracketing CircExons should be longer than those internally

the CircExon, so that would permit back splicing instead of normal splicing. As it is presented in Figure 3.15, internal introns are shorter (Table 3.6) than flanked introns in K562 cells.

Introns	**Mean**	**Median**	**Standard deviation**
Internal	5416.61	1915.5	14287.04
Bracketed	16461.82	5831	34416.92

Table 3.6: Mean, median and standard deviation of internal and bracketed introns.

3.7 GC content follow the same motif in two different cell lines.

So far, all analysis concerned the cell line K562, which is immortalised myelogenous leukemia cells and thus, carry characteristics of hematopoietic cells. In order to investigate wether the present patterns are cell specific or represented by more tissues in the human, the cell line SH-SY5Y was introduced in the research. This cell line is originated by human bone marrow cells but are widely used in scientific research for neuronal function and differentiation.

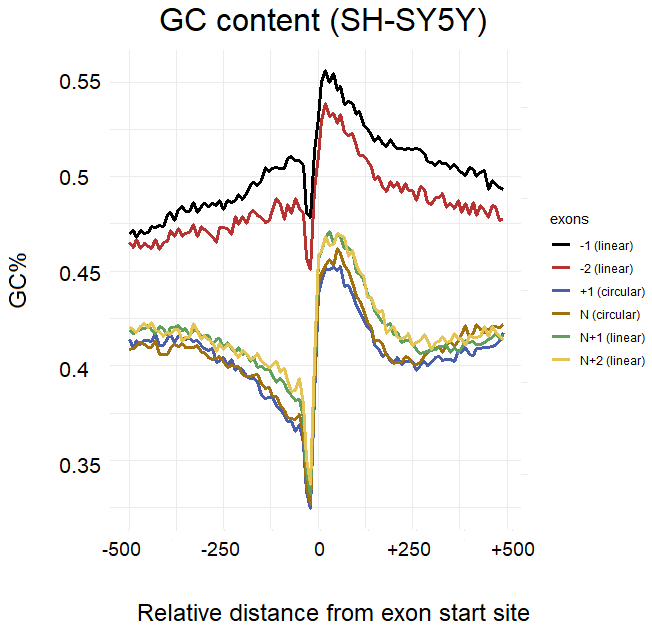


Figure 3.16: The neighborhood of the CircExon (GC content).

The patterns resulted from SymCurv prediction model, matched with those corresponding to K562 cell line (Figure 3.17). Again, the upstream linear exons have higher scores than the CircExon and the downstream linear exons, along to a rapid increment before ESS. Additionally, in Figure 3.16 and Table 3.7 it can be obsereved, that exons upstream +1 exon have more abrupt transitions to the GC content than -1 exons.

Exon Type	# exons	max Value	min Value
-2 (linear)	5750	0.54	0.45
-1 (linear)	9825	0.56	0.48
+1 (circular)	10471	0.45	0.32
+N (circular)	9825	0.46	0.33
N+1 (linear)	9081	0.47	0.33
N+2 (linear)	6767	0.47	0.34

Table 3.7: GC content minimum and maximum values for each exon categories.

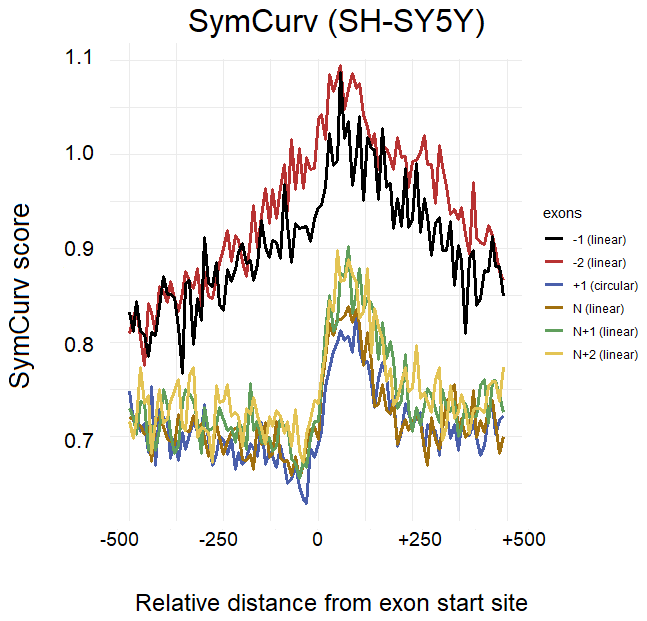


Figure 3.17: The neighborhood of the CircExon (SymCurv score).

Even though these two cell lines are used for different scientific purposes, they originally come from the bone marrow and thus, it is possible they express common transcripts which might give rise to CircExons. Indeed, only a very small percentage of the producted CircExons are tissue specific for SH-SY5Y (1%, compared with the K562 CircExons).

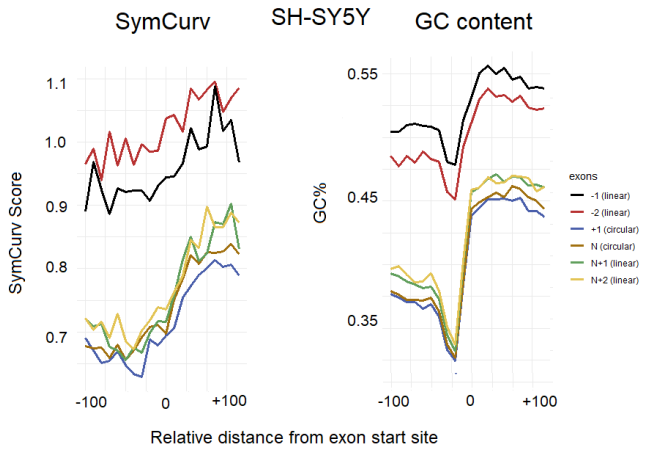


Figure 3.18: The neighborhood of the CircExon (ESS SymCurv and GC content comparison).

3.8 Cerebellum.

In the previous section, it was shown that CircExons expressed in endodermic cell lines have extremely similar consistence patterns (GC content and SymCurv prediction). Here, a dataset from CircExons expressed in cerebellum was downloaded from CIRCpedia, in order to investigate the patterns in cells form different germ layer. Cerebellum evaluated as a proper candidate, while neuron cells, originate from ectoderm. In contrast, bone marrow cells (K562 and SH-SY5Y) are endodermic. Additionaly, plenty studies of the recent years have demonstrated that circRNA expression is enriched in the brain, especially in the cerebellum (Gokoolparsadh et al., 2018; L. Li et al., 2017; You et al., 2015; Rybak-Wolf et al., 2015).

The first thing to investigate, was the GC content, while the access to MNase-seq data sets was not easy. Again, it was quite impressive to notice that the GC content pattern was similar with the content noticed in the previous sections for K562 and SH-SY5Y (Figure 3.23, Figure 3.10, Figure 3.16).

Symcurv analysis revealed the already proven motif for the ESS , leading with evidence of higher certainty that the pattern explaned in this work is universal representative for the circular exons of the human genome (Figure3.21, Table 3.9).

Furthermore a comparison was performed between K562 and SH-SY5Y cells and cerebellum, con-

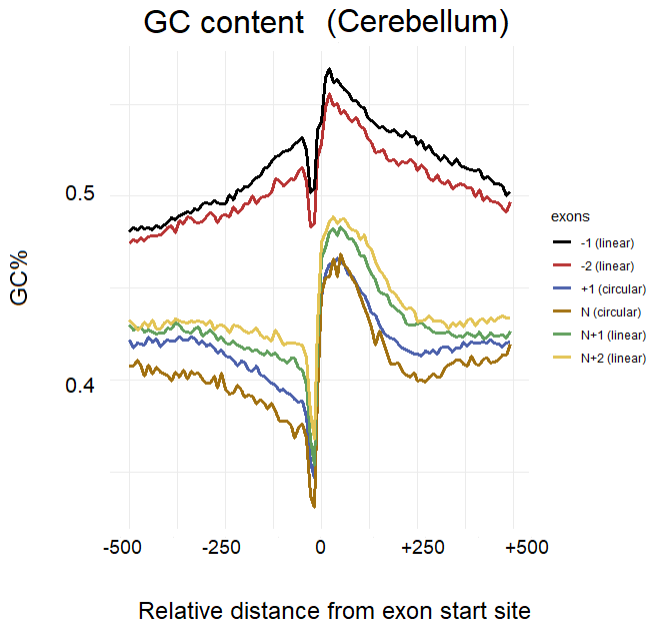


Figure 3.19: GC content around ESS of Cerebellum.

Exon Type	# exons	max Value	min Value
-2 (linear)	9033	0.56	0.47
-1 (linear)	16593	0.57	0.48
+1 (circular)	18536	0.47	0.35
+N (circular)	17413	0.46	0.33
N+1 (linear)	15003	0.48	0.35
N+2 (linear)	9707	0.49	0.37

Table 3.8: GC content minimum and maximum values for each exon categories.

cerning the characterization of the CircExons. It is important to mention that a unique exon, could be found in different relative positions in the neighborhood of the CircExon, while different transcripts of the same gene could give rise to different CircExons. For example, the exon SDF4_1159211_1159348 is found in 4 different transcripts and could be located in the internal of the CircExon, as well as in the initiation of the CircExon.

In Figure 3.20 are depicted the Circular Exons (CEi) of K562 cells and Cerebellum. Interestingly,

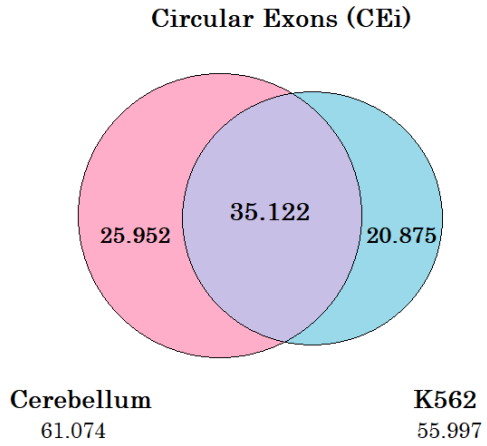


Figure 3.20: Cerebellum Linear Exons appear as Circular Exons in K562 cells.

57.5% of the Cerebellar CEi are common with the K562 CEi. The rest 42.5% of Cerebellar CEi, are expressed as linear exons in K562 or they are not expressed in the specific cell line.

he expression of CircExons is rather non-tissue specific. CIRCpedia contains 182.209 annotations of identified CircExons in cell lines and human organs. Each entry contains on average 14.068 CircExons but cerebrum express cumulatively, more CircExons than the other organs and cell lines. Three different cell types were analysed K562, SH-SY5Y and Cerebellum. Interestingly, K562 express 227 tissue specific CircExons, while Cerebellum express only 81.

Exon Type	**max Value**	**min Value**
-2 (linear)	1.13	0.83
-1 (linear)	1.11	0.81
+1 (circular)	0.87	0.66
N (circular)	0.88	0.69
N+1 (linear)	0.92	0.70
N+2 (linear)	0.96	0.72

Table 3.9: Symcurv minimum and maximum values for each exon category.

Finally, the depiction of transcript length giving rise to CircExons in Cerebellum, as described in

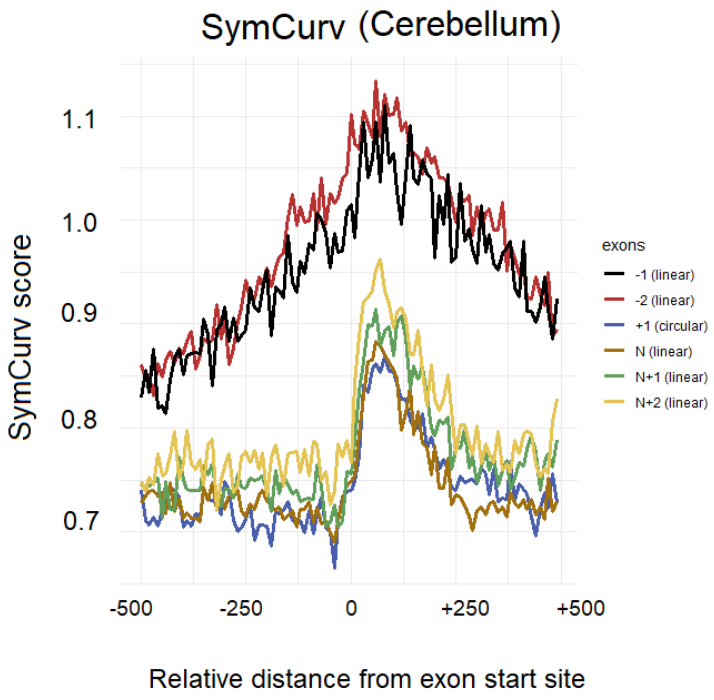


Figure 3.21: Symcurv analysis for nucleosome positioning prediction in Cerebellar CircExons.

Section 3.5 for K562, does not shown any important difference (Figure 3.22).

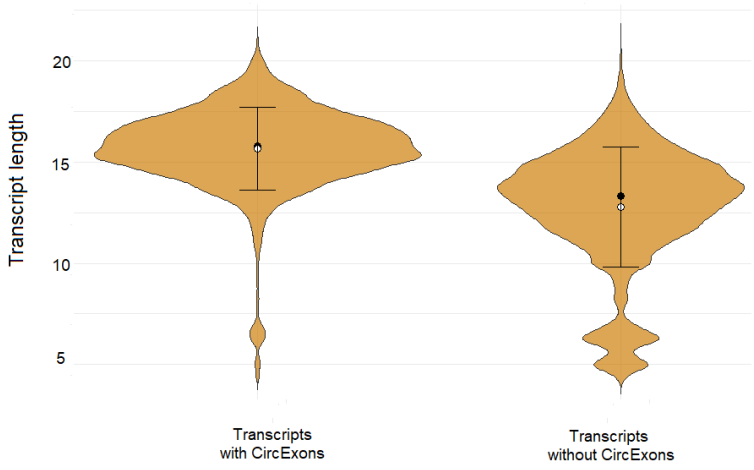


Figure 3.22: Transcripts giving rise to CircExons in Cerebellum tend to be longer than the average transcripts. All values are normilized with log2.

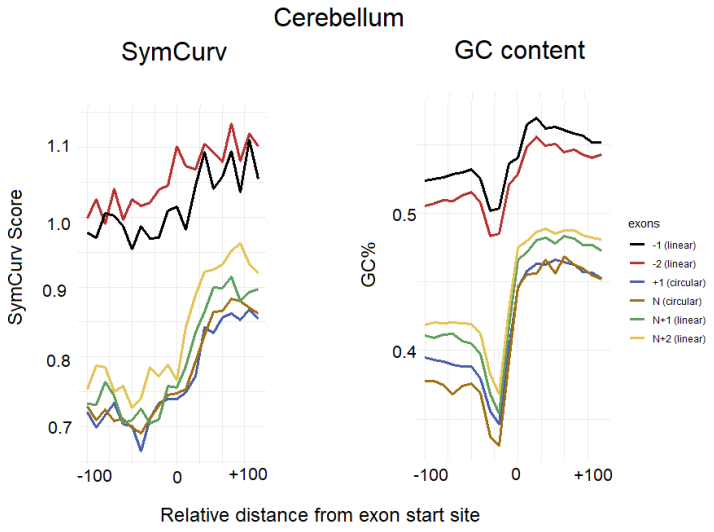


Figure 3.23: GC content around ESS of Cerebellum and SymCurv prediction.

Chapter 4

Discussion

The aim of the study was to reveal universal motifs, that represent human CircExons and differentiates them from the linear exons, in the context of the primary structure of the DNA, as well as in the context of the nucleosome organization around their boundaries. The most important result of the study is the GC content of CircExons, which has a unique motif in comparison with linear exons. However, when comparing CircExons with the downstream and upstream linear exons, CircExon GC content is strongly similar with the GC content of the downstream linear exons. On the other hand, upstream linear exons show a completely different motif which could be related to back splicing signaling. Those results consist new data in the research of circular RNAs and thus, further analysis in order to validate is recommended.

Additionally to the already presented results, when plotting the normalized GC% values (z-scores) of the circular and linear exons, it is clear that the GC content show a reduction 200bp upstream the ESS of the circular exons (Figure 4.1,4.2), along with a sharp dip. Those two consist important characteristics that CircExons own, and could contribute to back splicing signaling.

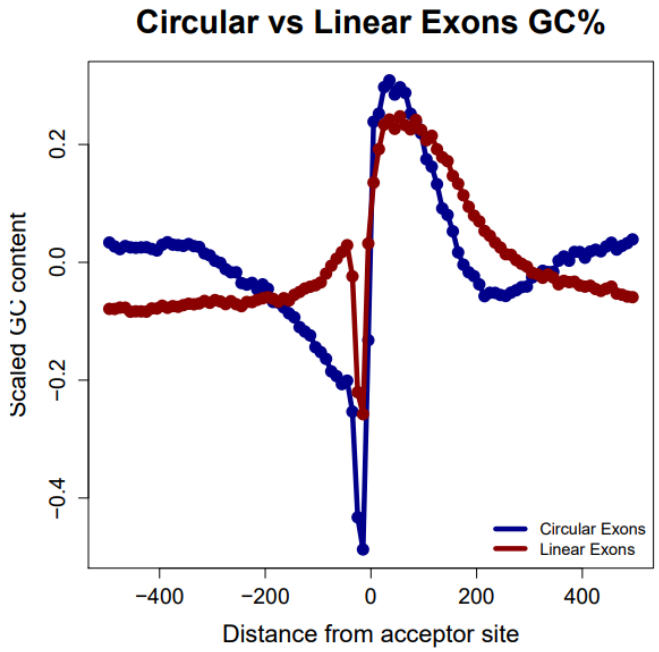


Figure 4.1: z-scores of GC content.

During the analysis, it was observed that many CircExons from CIRCpedia are consist of more than one exons and in particular, some CircExons' starts are not represented by an annotated ESS.

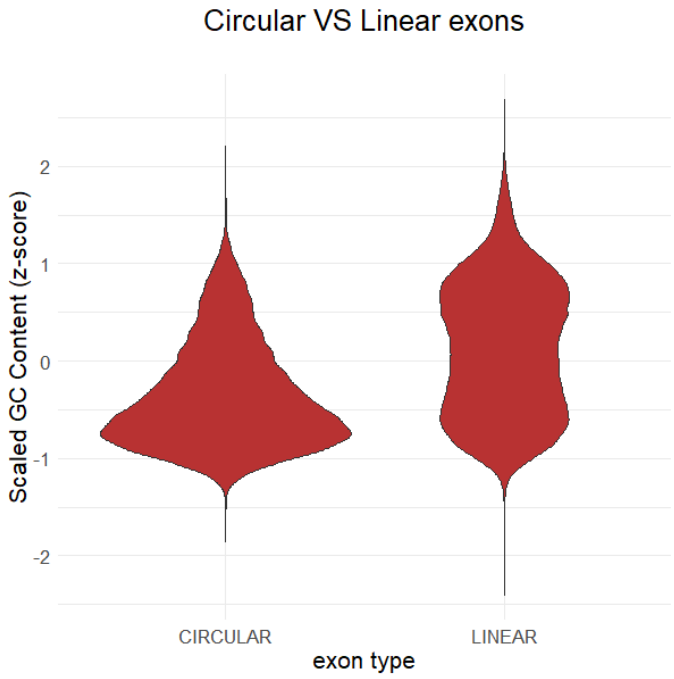


Figure 4.2: z-scores of GC content (200bp upstream ESS).

Similarly, some CircExons' end site are not annotated EES. This leads to the conclusion that as in alternative splicing, alternative 5'donor and 3'acceptor sites exist, it is possible that alternative 3'donor and 5' acceptor sites exist in back splicing. Thus, we consider back splicing as an "opposite" alternative splicing, which includes alternative 3'donor and 5' acceptor sites.

Taking under consideration the fact that the majority of the last exons of the CircExons (N) are not the terminal exon of the transcript (only 14,84% of the N exons correspond to the last exon of the parental transcript), it was important to investigate the downstream splice donor site of the circular exons, compared with these of the linear exons. The results proved that no significant differences are observed between downstream linear exons and the last exon of the CircExon (N), while the upstream linear exons have sharp GC content dips, which could be responsible for the loose nucleosome positioning, resulted by the MNase-seq analysis in K562 cells.

It is interesting that according to our findings, downstream linear exons have similar GC content motifs with circular exons, rather than with upstream exons. The first scenario we support, relies on the signaling role that the upstream linear exons have. We hypothesize that the content of the

upstream linear exons, is more crucial than the content of the downstream linear exons, while contributes to the recognition of the 5' back splice acceptor site. The second scenario refers to the nature of the downstream linear exons. It is possible that these exons are part of a CircExon, and thus carry characterists of CircExons in other cell types or are produced during different developmental processes (e.g cell differentiation or embryogenesis). They could also be unannotated CircExons.

As we presented above, SymCurv and MNase-seq signals are in agreement when linear exons were analysed (Figure 3.8) but not for exons that give rise to CircExons. This indicates a SymCurv weakness to predict the nucleosomic occupancy of CircExons. While Symcurv is strongly depended on the primal structure of DNA and thus, is depended on the GC content, it is clear that CircExon's strong nuclesomic signal, is related to additional determinant factors. When the two charts were scaled (Figure 4.3), it was impressive to note that Symcurv could not detect the long tailing of the CE nucleosomic signal. Both in linear and circular exons, SymCurv could not predict the sharp dip, which is present upstream of every ESS. Thus, we suggest that the mounted nucleosomes around the ESS are rather depended on other factors like the geometry of the molecule or the accessibility.

Additionally, it is shown that CircExons have stronger MNase-seq signal than the average and linear exon (Figure 3.4), but the tailing of CircExon curve, both in Mnase-seq and SymCurv is stronger than the tailing of linear exons. This indicates again, that nucleosome positioning is not depended exclusively on the sequence, even though sequence constitutes the fundamental determinant of molecular interactions. We suggest that this strong nucleosome positioning should determin back-splicing, while nucleosome existence is suggestive of splicing signals. One possible theory we support for the incompatibility between MNase-seq and SymCurv is that even though circular exons are expected to have lower occupancy, due to lower GC content, the nucleosome is blocked on the ESS, probably due to the existence of repetitive elements of high complementarity and thus the signal from the MNase-seq is stronger than in linear exons.

Another important finding resulted from this work, was that CircExons do not present tissue specific motifs. This was expected while most CircExons are common in every cell line we investigated. However we consider that common CircExons should exist in different expression levels in every cell line. It would be interesting to identify the expression of tissue specific transcripts giving rise to CircExons, in order to identify the regulatory role of these molecules.

Other groups (Jeck et al., 2013; Wilusz, 2015) have shown that repetitive elements of high complementarity, as sequences derived from transposons, flank exons. The effect of these sequences, relies

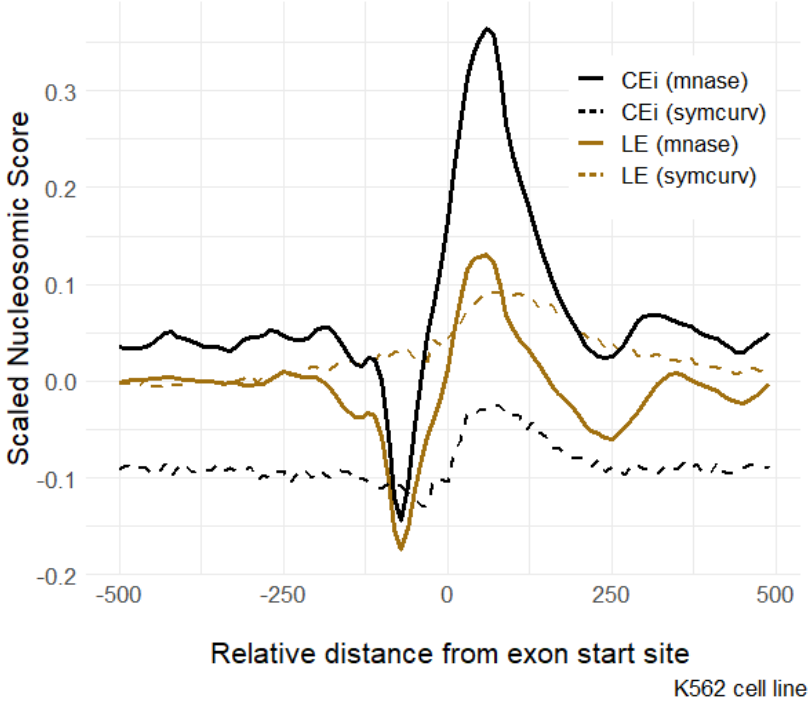


Figure 4.3: Comparisson between MNase-seq and SymCurv signal for linear and circular exons.

on the formation of an mRNA loop, with the splice sites into close proximity and thus, back splicing and circularization of the exon (or exons) are favored (Figure 4.4). Either way, back splicing and linear splicing require spliceosome assembly (Starke et al., 2015). Thus, in the context of analysing flanking CircExon sequences, GC content analysis of these elements would offer additional information for the determination of CircExon boundaries .

Unpublished data from our scientific team, show that CircExons are enriched in specific subcompartments of the chromatin. It is known that, A compartments are gene rich and have high GC content (Lieberman-Aiden et al., 2009). However, A2 subcompratment contains more CircExons than the other subcompartments of A. This is expected at first place, while A2 subcompartment contains longer transcripts than the others and CircExons, as shown in the present study, originate from long parental transcripts. Thus, a question worth been answered, is why CircExons are distributed in every compartment, while the characteristics are more likely to fit in the A2 subcompartment.

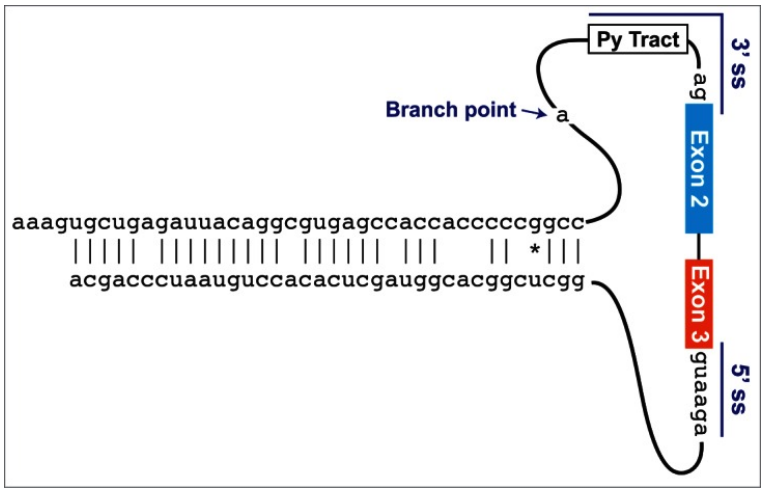


Figure 4.4: Alu elements are highly complementary to one another and can promote back splicing, by forming a hairpin. Source: Wilusz, 2015.

Taken together our data, we provide a universal characterization of CircExons in human cells, which contributes in the deeper comprehension of the mechanism that CircExons are produced. Thus, we expect that in the future, the impact of CircExon-mediated regulation will be clarified, while numerous publications mention the involvement of CircExons (and circRNAs in general) in cancer (Okholm et al., 2020; Starke et al., 2015; Lei, Tian, Fan, & Ni, 2019), as well as in human brain (Gokoolparsadh et al., 2018).

Bibliography

Aeling, K. A., Steffen, N. R., Johnson, M., Hatfield, G. W., Lathrop, R. H., & Senear, D. F. (2007). Dna deformation energy as an indirect recognition mechanism in protein-dna interactions. *IEEE/ACM transactions on computational biology and bioinformatics*, *4*(1), 117–125.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Keith Roberts, P. W., et al. (2018). Molecular biology of the cell.

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., & Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome research*, *19*(10), 1732–1741.

Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., ... Kadener, S. (2014). circrna biogenesis competes with pre-mrna splicing. *Molecular cell*, *56*(1), 55–66.

Bachmayr-Heyda, A., Reiner, A. T., Auer, K., Sukhbaatar, N., Aust, S., Bachleitner-Hofmann, T., ... Pils, D. (2015). Correlation of circular rna abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis and normal human tissues. *Scientific reports*, *5*(1), 1–10.

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *cell*, *136*(2), 215–233.

Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2001). Splicedb: database of canonical and non-canonical mammalian splice sites. *Nucleic acids research*, *29*(1), 255–259.

Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., ... Lovell-Badge, R.

(1993). Circular transcripts of the testis-determining gene sry in adult mouse testis. *Cell*, *73*(5), 1019–1030.

Chubb, J. R., & Bickmore, W. A. (2003). Considering nuclear compartmentalization in the light of nuclear dynamics. *Cell*, *112*(4), 403–406.

Clark, D. J., & Felsenfeld, G. (1992). A nucleosome core is transferred out of the path of a transcribing polymerase. *Cell*, *71*(1), 11–22.

Denisov, D., Shpigelman, E., & Trifonov, E. (1997). Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene*, *205*(1-2), 145–149.

Dong, R., Ma, X.-K., Li, G.-W., & Yang, L. (2018). Circpedia v2: an updated database for comprehensive circular rna annotation and expression comparison. *Genomics, proteomics & bioinformatics*, *16*(4), 226–233.

Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S.-p., Baldi, P. F., & Hertel, K. J. (2005). The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences*, *102*(45), 16176–16181.

Gokoolparsadh, A., Anwar, F., & Voineagu, I. (2018). The landscape of circular rna expression in the human brain. *bioRxiv*, 500991.

Guo, J. U., Agarwal, V., Guo, H., & Bartel, D. P. (2014). Expanded identification and characterization of mammalian circular rnas. *Genome biology*, *15*(7), 409.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., & Kjems, J. (2013). Natural rna circles function as efficient microrna sponges. *Nature*, *495*(7441), 384–388.

Hansen, T. B., Kjems, J., & Damgaard, C. K. (2013). Circular rna and mir-7 in cancer. *Cancer research*, *73*(18), 5609–5612.

Hennig, W. (1999). Heterochromatin. *Chromosoma*, *108*(1), 1–9.

Iannone, C., Pohl, A., Papasaikas, P., Soronellas, D., Vicent, G. P., Beato, M., & Valcárcel, J. (2015). Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. *Rna*, *21*(3), 360–374.

IIDA, Y., & SASAKI, F. (1983). Recognition patterns for exon-intron junctions in higher organisms as revealed by a computer search. *The Journal of Biochemistry*, *94*(6), 1731–1738.

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., ... Sharpless, N. E. (2013). Circular rnas are abundant, conserved, and associated with alu repeats. *Rna*, *19*(2), 141–157.

Kelly, S., Greenman, C., Cook, P. R., & Papantonis, A. (2015). Exon skipping is correlated with exon circularization. *Journal of molecular biology*, *427*(15), 2414–2417.

Kharchenko, P. V., Woo, C. J., Tolstorukov, M. Y., Kingston, R. E., & Park, P. J. (2008). Nucleosome positioning in human hox gene clusters. *Genome Research*, *18*(10), 1554–1561.

Kitamura-Abe, S., Itoh, H., Washio, T., Tsutsumi, A., & Tomita, M. (2004). Characterization of the splice sites in gt–ag and gc–ag introns in higher eukaryotes using full-length cdnas. *Journal of bioinformatics and computational biology*, *2*(02), 309–331.

Kogan, S., & Trifonov, E. N. (2005). Gene splice sites correlate with nucleosome positions. *Gene*, *352*, 57–62.

Kujirai, T., & Kurumizaka, H. (2020). Transcription through the nucleosome. *Current Opinion in Structural Biology*, *61*, 42–49.

Kwek, K. Y., Murphy, S., Furger, A., Thomas, B., O'Gorman, W., Kimura, H., ... Akoulitchev, A. (2002). U1 snrna associates with tfiih and regulates transcriptional initiation. *Nature structural biology*, *9*(11), 800–805.

Lai, W. K., & Pugh, B. F. (2017). Understanding nucleosome dynamics and their links to gene expression and dna replication. *Nature reviews Molecular cell biology*, *18*(9), 548.

Lei, B., Tian, Z., Fan, W., & Ni, B. (2019). Circular rna: a novel biomarker and therapeutic target for human cancers. *International journal of medical sciences*, *16*(2), 292.

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–719.

Li, L., Zheng, Y.-C., Kayani, M. U. R., Xu, W., Wang, G.-Q., Sun, P., ... others (2017). Comprehensive analysis of circrna expression profiles in humans by raise. *International Journal of Oncology*, *51*(6), 1625–1638.

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., ... others (2015). Exon-intron circular rnas regulate transcription in the nucleus. *Nature structural & molecular biology*, *22*(3), 256.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A.,

... others (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, *326*(5950), 289–293.

Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature reviews Molecular cell biology*, *13*(7), 436–447.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, *389*(6648), 251–260.

Martins, S. B., Rino, J., Carvalho, T., Carvalho, C., Yoshida, M., Klose, J. M., ... Carmo-Fonseca, M. (2011). Spliceosome assembly is coupled to rna polymerase ii dynamics at the 3 end of human genes. *Nature structural & molecular biology*, *18*(10), 1115.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., ... others (2013). Circular rnas are a large class of animal rnas with regulatory potency. *Nature*, *495*(7441), 333–338.

Okholm, T. L. H., Sathe, S., Park, S. S., Kamstrup, A. B., Rasmussen, A. M., Shankar, A., ... others (2020). Transcriptome-wide profiles of circular rna and rna binding protein interactions reveal effects on circular rna biogenesis and cancer pathway expression. *BioRxiv*.

Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., ... others (2017). Translation of circrnas. *Molecular cell*, *66*(1), 9–21.

Passarge, E. (1979). Emil heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. *American journal of human genetics*, *31*(2), 106.

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and gc content of the human genome. *BMC research notes*, *12*(1), 106.

Roy, M., Kim, N., Xing, Y., & Lee, C. (2008). The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *Rna*, *14*(11), 2261–2273.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., ... others (2015). Circular rnas in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Molecular cell*, *58*(5), 870–885.

Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., & Kleinschmidt, A. K. (1976). Viroids are single-stranded covalently closed circular rna molecules existing as highly base-paired rod-like structures. *Proceedings of the National Academy of Sciences*, *73*(11), 3852–3856.

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core dna. *Journal of molecular biology*, *191*(4), 659–675.

Schwartz, S., Meshorer, E., & Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature structural & molecular biology*, *16*(9), 990.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., ... Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, *442*(7104), 772–778.

Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.-H., & Bindereif, A. (2015). Exon circularization requires canonical splice signals. *Cell reports*, *10*(1), 103–111.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., & Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature structural & molecular biology*, *16*(9), 996.

Tillo, D., & Hughes, T. R. (2009). G+ c content dominates intrinsic nucleosome occupancy. *BMC bioinformatics*, *10*(1), 442.

Venkatesh, S., & Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. *Nature reviews Molecular cell biology*, *16*(3), 178–189.

Wang, Y., Liu, J., Huang, B., Xu, Y.-M., Li, J., Huang, L.-F., ... others (2015). Mechanism of alternative splicing and its regulation. *Biomedical reports*, *3*(2), 152–158.

Weintraub, H., & Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. *Science*, *193*(4256), 848–856.

Wilusz, J. E. (2015). Repetitive elements regulate circular rna biogenesis. *Mobile genetic elements*, *5*(3), 39–45.

Wolffe, A. (1998). *Chromatin: structure and function*. Academic press.

You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., ... others (2015). Neural circular rnas are derived from synaptic genes and regulated by development and plasticity. *Nature neuroscience*, *18*(4), 603–610.

Zhang, Y., Xue, W., Li, X., Zhang, J., Chen, S., Zhang, J.-L., ... Chen, L.-L. (2016). The

biogenesis of nascent circular rnas. *Cell reports*, *15*(3), 611–624.

Zhao, X., Cai, Y., & Xu, J. (2019). Circular rnas: biogenesis, mechanism, and function in human cancers. *International journal of molecular sciences*, *20*(16), 3926.