UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

LUÍZA ZUVANOV DE FARIA

Study of evolution and architecture of minimal introns

São Carlos
2020

LUÍZA ZUVANOV DE FARIA

Study of evolution and architecture of minimal introns

Dissertation presented to the Graduate Program in Physics at the Instituto de Física de São Carlos, Universidade de São Paulo to obtain the degree of Master of Science.

Concentration area: Biomolecular Physics
Advisor: Prof. Dr. Ricardo De Marco

Corrected Version

(original version available on the Program Unit)

São Carlos

2020

Zuvanov, Luíza
Study of evolution and architecture of minimal introns / Luíza Zuvanov; advisor Ricardo De Marco - corrected version -- São Carlos 2020.
164 p.

Dissertation (Master's degree - Graduate Program in Biomolecular Physics) -- Instituto de Física de São Carlos, Universidade de São Paulo - Brasil , 2020.

1. Minimal intron. 2. Homeothermy. 3. Cell division. 4. Intron retention. 5. Spermatogenesis. I. De Marco, Ricardo, advisor. II. Title.

To my daughter, Anna.

ACKNOWLEDGEMENTS

I would like to pay my special regards to Prof. Ricardo De Marco, my supervisor, who made this scientific work a joyful and enriching moment. He gave me the opportunity for discussing our hypotheses, discarding some, elaborating new ones, but always keeping the curiosity and passion for science.

I also want to express my gratitude for all professors and colleagues that guided me to the scientific career. I would like to thank Prof. Daniela Mara de Oliveira, who was my first supervisor during my degree in Biological Sciences, and my laboratory colleague and friend M.Sc. Paula Queiroz Alvim of the Mesenchymal Stem Cell group from the University of Brasília, where I gave my first steps in the scientific journey. To Dr. Alain Kohl and Dr. Claire Donald from the Centre for Virus Research at the University of Glasgow for all the shared knowledge. To Prof. Antônio Francisco Pereira de Araújo who supervised me in my first project in bioinformatics and taught me precious knowledge about the information theory. I would like to give a special thanks to Prof. Ilana Camargo, who first received me in the Institute of Physics of São Carlos and taught me some valuable lessons about organization and discipline. I also want to thank the members of the Molecular Biology and Genome Evolution Group from the Institute of Physics of São Carlos for the discussions on bioinformatics and evolution.

I wish to express my deepest gratitude to Prof. Bergmann Morais Ribeiro who gave me a great opportunity to develop my scientific skills during the years in his laboratory. A special thanks to my former supervisor and friend Dr. Fernando Lucas de Melo who was one of my most important influences when choosing the scientific career. To all the team of the Baculovirus group from the University of Brasília, in special, Dr. Fabricio Morgado who supervised me and taught me some clever solutions to experimental problems.

I would like to recognize the importance of the National Council for Scientific and Technological Development (CNPq) that allowed me to dive into the science world by financially supporting me during my participation in the programs for scientific initiation, exchange program at the University of Glasgow and my master's degree. I

also like to thank all the staff from the Institute of Physics who also made this work possible. Without this support, I could not do science.

I would like to offer my special thanks to all my friends and family who have made these years happier. In special, to my friends Bárbara de Luca and Laís Ribeiro for all the talks and laughs. To my lab mates for daily support and friendship. To my grandparents and stepmother. My deepest gratitude to my parents Adriana Sampaio Zuvanov and Weber Lemes de Faria who raised me with all their love and encouraged me to be curious about the world that brought me to where I am today. To my beautiful sisters Luana Zuvanov de Faria and Júlia Perini de Faria for being my confidants and my best friends. And finally, to my beloved fiancé André Marcos Perez for always standing by me, for making these past years the greatest years of my life. To Angus and Trufa for their fluffiness.

“Nothing in biology makes sense except in the light of evolution.”

Theodosius Dobzhansky (1973)

ABSTRACT

ZUVANOV, L. **Study of evolution and architecture of minimal introns.** 2020. 164p. Dissertation (Master in Science) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2020.

Eukaryotic introns show a wide span of size, from only 30bp to large 3.6Mbp. However, analysis of intron size distribution in diverse lineages shows a frequent accumulation of introns near the minimum size that is referred to as minimal introns. In this work, structure and evolution of minimal introns were studied based on diverse species of bilaterian animals, especially the platyhelminth *Schistosoma mansoni* and species from the Vertebrata phylum. Analysis of the distribution of introns size from *Schistosoma mansoni* shows a minimal intron peak at 34bp, a remarkably short size when compared to other eukaryotic species. Minimal introns from *Schistosoma mansoni* were preferentially found in some specific chromosomes. While studying intron retention (IR) and splicing signals, it was observed that premature termination codons (PTC) were preferentially found in the second and last codons of minimal introns due to contribution of the splice sites sequences. Symmetric minimal introns display the highest proportion of PTC-containing introns. We speculate that this observation reflects an evolutionary pressure associated with the fact that its retention does not shift the reading frame of translation. Interestingly, the proportion of PTC-containing introns does not increase with size in symmetric minimal introns as observed for non-symmetric minimal introns. We suggest that a large fraction of symmetric minimal introns that do not present PTC may be retained for the production of isoforms with few additional amino acid residues. The lack of preference of minimal introns with PTC for any position along the gene suggests that nonsense-mediated decay of *Schistosoma mansoni* is independent of exon junction complex (EJC). Study of the evolution of minimal introns from vertebrates shows that the acquisition of homeothermy had a great influence on minimal introns GC%. In high body temperature species, minimal introns can be divided into low and high GC% populations, with peaks of ~30% and ~70% respectively. Analysis of the human genome shows that, although the GC% variation was more prominent in minimal introns, the entire gene sequence varies. Genes without minimal introns do not appear to show GC% variation dependent on the body temperature. This suggests that minimal introns can serve as proxies for

detecting temperature-responsive genes in humans. The transition from low to high GC% of some minimal introns may be impaired due to high IR levels in the intermediate GC%. Low GC% minimal intron-containing genes were related to cell division and thus transition to high GC% may be cumbersome due to high levels of IR. Furthermore, genes with low GC% minimal introns were observed to be related to oncogenic transformation and to be highly expressed in the meiosis process. Based on these results, we propose that minimal intron-containing genes could represent a new interesting system for studying diseases related to division defects, such as cancer and infertility. Also, as IR has been observed to be an important factor for selecting minimal introns GC%, the participation of minimal intron-containing genes in diseases in which increase of IR could be associated, such as diabetes type I and cancer, may be further explored.

Keywords: Minimal intron. Homeothermy. Cell division. Intron retention. Spermatogenesis.

RESUMO

ZUVANOV, L. **Estudo da evolução e arquitetura de íntrons mínimos.** 2020. 164p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2020.

Íntrons de eucariotos possuem diversos tamanhos, desde diminutos 30pb a 3.6Mpb. Entretanto, análises da distribuição de tamanho de íntrons em diversas linhagens mostram um frequente acúmulo de íntrons próximos ao tamanho mínimo, os quais são chamados de íntrons mínimos. Neste trabalho, a estrutura e evolução de íntrons mínimos foram estudadas em diversas espécies de animais bilaterais, especialmente o platelminto *Schistosoma mansoni* e espécies do filo Vertebrata. Análise de distribuição de tamanho de íntrons de *Schistosoma mansoni* mostra um pico de íntrons mínimos em 34pb, um tamanho notavelmente pequeno se comparado ao de outras espécies de eucariotos. Íntrons mínimos de *Schistosoma mansoni* foram preferencialmente encontrados em certos cromossomos. Ao estudar retenção intrônica (IR) e sinais de *splicing*, observou-se que códons de parada prematura (PTC) são preferencialmente encontrados no segundo e último códons de íntrons mínimos devido à contribuição das sequências dos sítios de *splicing*. Íntrons mínimos simétricos apresentam maior proporção de íntrons com PTC. Sugerimos que essa constatação reflita pressões evolutivas associadas ao fato da retenção destes íntrons não mudarem o quadro de leitura de tradução. Interessantemente, a proporção de íntron mínimos com PTC não aumenta conforme o tamanho de íntrons mínimos simétricos como visto em íntrons mínimos não simétricos. Sugerimos que grande parte de íntrons mínimos simétricos que não possuam PTC possam ser retidos para produção de isoformas com alguns resíduos de aminoácidos adicionais. A não preferência de íntrons mínimos com PTC por posições ao longo do gene sugere que o decaimento mediado por códons de parada (NMD) independe do complexo de junção de éxons (EJC). Estudo da evolução de íntrons mínimos de vertebrados mostra que a aquisição da homeotermia teve grande influência no conteúdo GC de íntrons mínimos. Em espécies com alta temperatura corpórea, os íntrons mínimos podem ser divididos em populações de baixo e alto GC%, com picos de ~30% e ~70% respectivamente. Análises do genoma humano mostram que, embora variações de GC% sejam proeminentes em íntrons mínimos, sequências gênicas inteiras variam.

Genes sem íntrons mínimos não apresentam aparente variação de GC% dependente de temperatura corpórea. Isso sugere que íntrons mínimos possam servir como marca para identificação de genes responsivos à temperatura em humanos. A transição de baixo para alto GC% por alguns íntrons mínimos pode ser comprometida devido aos altos níveis de retenção no GC% intermediário. Genes com íntrons mínimos de baixo GC% são relacionados a divisão celular e, portanto, a transição para alto GC% pode ser complicada devido aos altos níveis de retenção. Ademais, observou-se que genes com íntrons mínimos de baixo GC% estão relacionados à transformação oncogênica e que são altamente expressos na meiose. Baseando-se nestes resultados, propomos que genes com íntron mínimo possam representar um novo sistema de estudo de doenças relacionadas a defeitos de divisão celular, como câncer e infertilidade. Ainda, devido a IR ser um fator importante para seleção de GC% de íntrons mínimos, deve-se melhor explorar a participação de genes com íntrons mínimos em doenças nas quais o aumento da IR possa estar associado, como diabetes tipo I e câncer.

Palavras-chave: Íntron mínimo. Homeotermia. Divisão celular. Retenção intrônica. Espermatogênese.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

3n	Size multiple of 3
A	Adenine
ATP	Adenosine triphosphate
bp	Base pair unit
C	Cytosine
CDS	Coding sequence
DNA	Deoxyribonucleic acid
EJC	Exon junction complex
G	Guanine
HSPA2	Heat shock protein A2
IEP	Intron encoded protein
IR	Intron retention
KDE	Kernel density estimation
Lg	Logarithm of base 10
mBBP	Mammalian branch point binding protein
mRNA	Messenger ribonucleic acid
N	Adenine / Guanine / Cytosine / Thymine
n	Sample size
NMD	Nonsense-mediated decay
nt	Nucleotide unit
PIR	Percentage of intron retention
PTC	Premature termination codon
R	Guanine / Adenine (purine)
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SF1	Splicing factor 1
snRNA	Small nuclear ribonucleic acid
snRNP	Small nuclear ribonucleoprotein
sp	Species (singular)
spp	Species (plural)

T	Thymine
TPM	Transcripts per million
tRNA	Transfer ribonucleic acid
TS	Tissue score
U	Uracil
UTR	Untranslated region
Y	Cytosine / Thymine (pyrimidine)
τ	Tau, gene expression tissue-specificity metric

CONTENTS

Part I

INTRODUCTION TO INTRONS

Introns are entities that reside inside the gene and divide it into discrete units called exons. It is not surprising that introns were one of the most enigmatic discoveries in 20th century. [1] For long, introns were considered a puzzle to molecular biology and have put to question the flow of genetic information and some other fundamental concepts. In this first part of the dissertation, we expect the reader to have a brief introduction to some fundamentals aspects of nature and consequences involving the existence of introns. We start our study by discussing how introns were discovered and the characteristics that have divided them into four groups. We then highlight some evidence that might explain their evolutionary origins. In chapter 2, we give the reader some further details about one of those groups, the spliceosomal introns that have been essential for the understanding of the evolution of eukaryotes. We discuss the mechanism of the excision of the spliceosomal introns and the consequences of having them in the genome. Finally, we examine the intron retention phenomena (IR), a form of alternative splicing that has an important role in the regulation of RNA stability. In the last chapter of this section, we introduce our main object of study of this dissertation, the minimal introns, highlighting some of their interesting aspects. We hope that this pertinent introduction gives us a solid background that enables us to have valuable discussions and conclusions of our work.

Chapter 1

THE HISTORY OF INTRONS: DISCOVERY AND EVOLUTION

Introns, the non-coding sequences found inside a gene, were only discovered around 40 years ago, [2-3] but ever since have been considered a puzzle in molecular biology. The unexpected discovery of introns with uncertain functions and origins put into question if they were genomic parasites or even “junk DNA”. Their existence made scientists revise some basic principles in the flow of genetic information. Transcripts should be processed before proper translation. After introns discovery, the key questions “Why genes in pieces?” posed by Walter Gilbert or even “Were they ever together?” posed by Ford Doolittle have been answered with the effort of many scientists around the world. [4-5] One thing is for sure, introns have been involved with some important evolutionary implications for their host. The evolution of eukaryotes, for example, have an intrinsic relation to the consequences of harbouring introns. [1,6]

In this chapter, we proposed a brief introduction to these entities that have divided genes into discrete units. We start our study by discussing how introns were discovered and the implications of this discovery. Then, we summarize and describe the unique features of the four groups that introns are categorized. In the end, we succinctly discuss one of the fundamental hypotheses of the evolutionary origin of introns.

1.1 The discovery of introns

In the middle of 1970, genes were considered as continuous entities in the double-stranded helix of the DNA. However, in the year of 1977, Richard J. Roberts and Philip A. Sharp changed the commonly held view regarding the structure of the

molecule of heredity. While they were independently studying a common cold virus to determine the location of different genes in its DNA genome, they realised that a biochemical experiment was not working as expected. This experiment, called R loop, consists in hybridizing the mRNA with the melted template DNA for determining the exact position where the transcription occurred. In that particular case, they found that the 5′ end of the RNA wasn't located in the immediate vicinity of the rest of the gene. Observation of the hybridized double-strand under the electron microscope showed that the mRNA was discontinuously hybridized to the template DNA (Figure 1.1). Roberts and Sharp soon came to the conclusion that actually the gene was made up of several separated segments that earned them the Nobel Prize in Physiology or Medicine in 1993. [2-3,7]

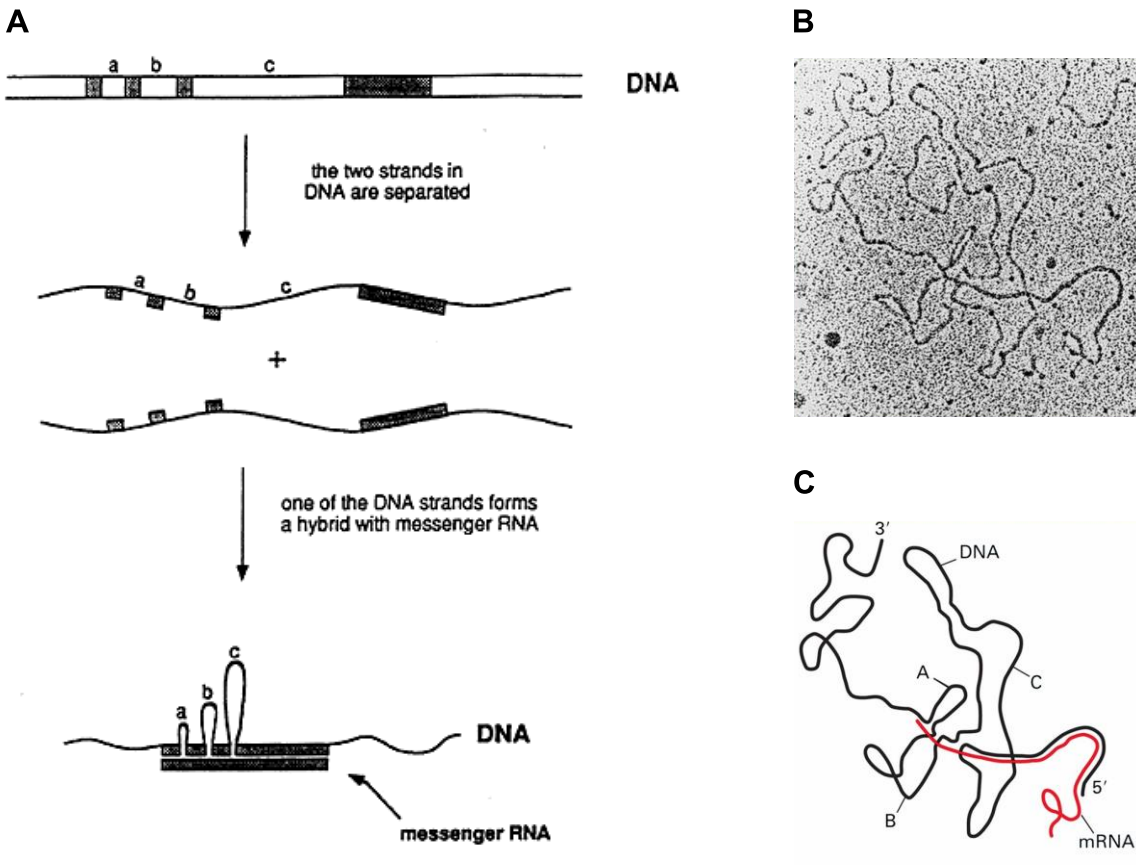


Figure 1.1 - Schematic representation of the R loop experiment used to discover the fragmented nature of the gene (a). The mRNA molecule is derived from four segments of the template DNA separated by three intervening sequences represented by letters a, b and c. The actual electron microscopy showing the segmented pairing of the mRNA molecule with its template strand of DNA is shown in (b) and its schematic interpretation in (c). The mRNA molecule is coloured in red and the DNA in black. The three intervening sequences are represented by the A, B and C unpaired segments in the loop format.

Source: (a) THE NOBEL PRIZE IN PHYSIOLOGY OR MEDICINE; [7] (b, c) BERK. [8]

The discovery that mRNA was confected by only some parts of the DNA led to the idea that the gene was, in fact, a combination of transcribing sequences interrupted by “silent” DNA (Figure 1.1). Some later analysis of diverse eukaryotic genes in many laboratories also predicted that genes were composed of separated segments. This observation, however, was only later detected in prokaryotic genomes. [9–11] As so, the considered “higher” organisms’ genes were a mosaic of expressed sequences, named exons, conjugated with non-coding sequences, known as introns as a reference of intragenic regions suggested by Walter Gilbert. In this context, the idea of the cistron necessarily corresponding to a continuous polypeptide chain was deprecated. As a consequence, it was predicted that regions corresponding to introns would be lost in the mature mRNA by an excision process known as splicing (Figure 1.2). This mechanism constitutes an important new step in the flow of genetic information. [4,7]

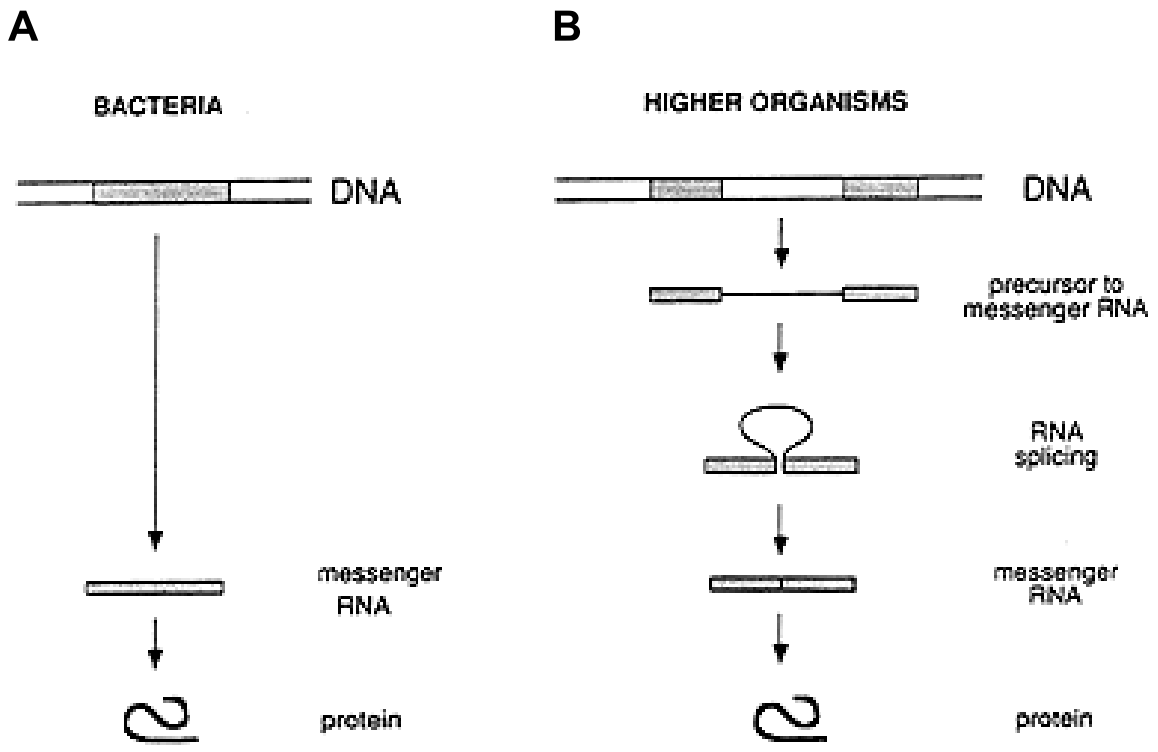


Figure 1.2 - Representation of the flow of the genetic information seen in bacteria (a) and in “higher” organisms (b). While in bacteria the mRNA is derived from a continuous segment of the DNA, in higher organisms, the information stored in the DNA is fragmented into exons. The intervening sequences, introns, found in the precursor to mRNA must be removed by the splicing process before being translated into protein.

Source: THE NOBEL PRIZE IN PHYSIOLOGY OR MEDICINE. [7]

1.2 Types of introns

Further studies allowed scientists to classify introns based on its excision mechanisms. There are four known groups of introns: group I, group II, tRNA, and spliceosomal introns. The groups I and II are called self-splicing introns, whereas the other two groups can only be spliced with the assistance of other specialized molecules. The distribution of introns into groups can be briefly summarized by some common features they might share, as described below:

- Group I: the introns classified in this group were detected in a variety of organisms, such as algae, fungi, lichens, lower eukaryotes and bacteria. They can be found in genes that code for proteins, rRNA or even tRNA. Its excision mechanism is based on its well-conserved secondary structure. Using an exogenous guanosine nucleotide attached to one of its helix structures, the 5′ end of the intron is nucleophilic attacked and forms a phosphodiester bond to the first nucleotide of the intron. Later, the free 3′-hydroxyl group from the upstream exon reacts with the 3′ splice site of the intron resulting in exons ligation and excision of the intron (Figure 1.3a). [12-13]

- Group II: this type of introns are ribozymes (catalytic RNA) and retroelements (elements capable of transposing themselves to other parts of the genome). They are found in the genomes of bacteria, archaebacteria, and some organelles of eukaryotes. Group II introns have a conserved secondary structure characterized by the presence of six domains. [14] Even though the splicing can over go with only the existence of Mg^{2+}, the process could be enhanced by the recruitment of host proteins. Basically, the excision and ligation processes are done in two steps: cleavage of the 5′ splice site by a nucleophilic attack, followed by the reaction between the 3′-hydroxyl in the upstream exon and the 3′ splice site (Figure 1.3b). The nucleophile used in the first step could either be an adenosine nucleotide found in one of the domains of the introns, or a water molecule. [15] Interestingly, some group II introns, instead of relying on host-proteins to improve their splicing process, are able to encode an intron

encoded protein (IEP). Typically, this IEP contains a reverse transcriptase domain that is used to facilitate intron splicing and mobility. [14,16]

- tRNA introns: those introns are present in all three domains of life, especially in archaeal and eukaryotic domains. tRNA introns are known to have their splicing process catalysed by a protein complex. The introns are removed by the action of an endonuclease and the exons are then ligated by a ligase enzyme (Figure 1.3c). Even though the splicing steps processes show unique features in archaeal and eukaryotic species by displaying different *cis* and *trans*-acting elements, almost all tRNA introns can be found inserted at a canonical position: one nucleotide after the anticodon. This characteristic makes the correct process of splicing crucial for tRNA maturation and appropriate function. [17-18]

- Spliceosomal introns: this class of introns are known for being one of the defining features of the eukaryotic lineage. [19] In fact, they can be found in the nuclear genomes of nearly all characterized eukaryotes. Whereas for the group I and group II approximately 1,500 and 200 cases have been characterized, respectively, hundreds of thousands of spliceosomal introns can be described in a typical genome of a higher eukaryote. Its splicing process is guided by the use of the spliceosome, a complex that comprises hundreds of proteins and five RNAs. [20] Even though those introns possess *quasi*-random sequences, they do share some conserved elements that are used as recognition marks by the spliceosome. In their 5′ end, the splice site is characterized by the presence of the GU dinucleotide, and in the other extremity of the intron, there is the branch point, followed by a polypyrimidine tract and a terminal AG 3′ splice site. The splicing process consists of two transesterification steps catalysed by the spliceosome complex. The first reaction occurs between the 5′ splice site and the adenine nucleotide found in the branch point that will create a lariat structure in the intron. Then, the free 3′-hydroxyl of the upstream exon reacts with the phosphate (P) of the 3′ splice site. These steps result in exon ligation and intron release in a lariat form (Figure 1.3d). [21]

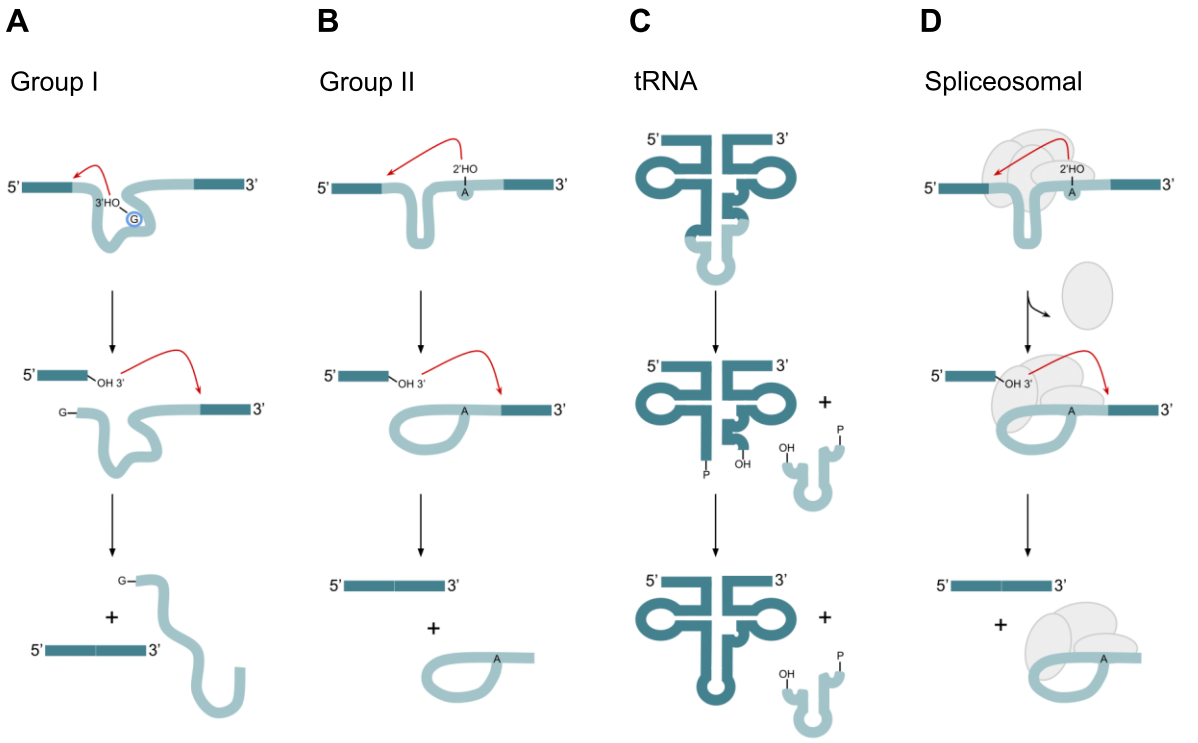


Figure 1.3 - Differences between the processing pathway of the four groups of introns. The splicing mechanism is characterized by the removal of introns and the ligation of exons through two consecutive transesterification reactions. Groups I (a) and II (b) are called self-splicing introns, whereas tRNA (c) and spliceosomal introns (d) need the assistance of a protein-only complex (not shown) and the spliceosome complex made of proteins and RNAs, respectively.
Source: (a, b, d) Adapted from WATSON; [22] (c) SCHMIDT; MATERA. [18]

1.3 The origin and evolution of introns

The history of where and when the different classes of introns first appeared and evolved are still being discovered. However, there are some good pieces of evidence that might explain their origins. Interestingly, this narrative may trace back to the “RNA world” era, which precedes biology based on DNA and protein molecules. In this context, groups I and II of self-splicing introns might have their origins on primordial ribozymes. [12,14]

The characteristics of group I introns suggest that they might have aroused from a molecule with autocatalytic properties which evolution seems to be intimately related to the theory of the “RNA world”. The well-established abilities of RNA molecules to store information and to promote catalysis suggest their capacity of self-replication

without the help of a protein. Indeed, it was described the potential of group I introns to polymerase RNA chains and to catalyse RNA ligation. Group I introns from *Anabaena* sp. can catalyse the formation of a phosphodiester bond with a mechanism that resembles the activity of the polymerase protein. [23] Furthermore, group I introns can move from its site to another in a transposition (in the same organism) or a horizontal gene transfer manner (from one organism to another) through reverse splicing. Those observations suggest that modern group I introns emerged when a primordial replicase-like ribozyme became capable of mobilization as a parasitic genetic element, that resembles the characteristics of this type of introns. [12]

Similarly to group I introns, the evolution of group II introns ribozyme portion might trace back to the “RNA world”. The phylogenetic distribution of group II introns in eubacteria and eukaryotic organelles suggests that their ancestor was like to be present in eubacteria before the emergence of eukaryotes and was transmitted to the eukaryotic last common ancestor through the endosymbiotic bacteria that give origin to mitochondria and chloroplast. This is because this type of introns is rare in archaebacteria, which are more closely related to the prokaryotic ancestor of eukaryotes than the eubacteria group. [16,24] The current understanding of group II introns origins is known as the “retroelement ancestor hypothesis”. In this model, the group II introns might have originated from a retroelement form with self-splicing and retromobility properties found in eubacteria. [14,16]

Contrary to these other intron classes, the spliceosomal introns are thought to appear later in the evolution. A variety of observations suggest that spliceosomal introns arose from ancestral group II introns. Given the facts that: group II introns are capable of replicative movements to new locations facilitated by their own reverse transcriptase, and the recurrent transfer of organelle genes to the nuclear genome, [25-26] it is suggested that the group II introns - transmitted to eukaryotes through the endosymbiosis hypothesis, have also evaded the eukaryotic nucleus and proliferated to many genomic sites. [16,27] The later recruitment of some nuclear proteins to the formation of a spliceosomal complex is also expected, as the group II introns are known for utilizing protein factors to improve their splicing. This leading group II-origin hypothesis has either support from structural and functional similarities of splicing between those two groups of introns and also from the phylogenetic distribution of

spliceosomal introns. This type of introns is found in every basal eukaryotic organism, but is absent in the genomes of prokaryotes, suggesting that spliceosomal introns have aroused soon after the emergence of the eukaryotic stem. [27]

Chapter 2

THE SPLICEOSOMAL INTRONS OF EUKARYOTES

In this chapter, a more detailed view of the role and mechanism of the spliceosomal introns will be presented. The understanding of spliceosomal introns is a matter of great interest in the study of the eukaryotic lineage and genome evolution. It is known that the splicing process is affected by many factors that could act antagonistically or synergistically across several degrees of complexity: from consensus sequences found in the border of introns to histone modifications at a chromatin level. To this extent, it will be discussed what features characterizes and differentiate an intron from an exon, making possible its recognition and successful removal. [28] Furthermore, the consequences of spliceosomal introns in the eukaryotic genome and their retention in the transcripts are also reviewed.

2.1 The splicing mechanism

The splicing process can be summarized as the mechanism in which an intron is removed and the adjacent exons are ligated in the RNA molecule. In eukaryotes, this process is catalysed by the spliceosome, one of the most complex machinery in the cell made of proteins and small nuclear ribonucleoproteins (snRNP) that orchestrates the binding and release of several hundred particles. The spliceosome is capable of promoting the excision of introns of different lengths and sequences. [28,29] There are only four known conserved sequences inside introns that are used for its recognition by the spliceosome: a 5′ and 3′ splice sites found in the border of the intron - commonly characterized by the presence of the GU and AG dinucleotides respectively; a branch point sequence found approximately in 18-40 nucleotides

upstream the 3′ splice site; and also a polypyrimidine tract between the branch point sequence and the 3′ splice site that is found in the metazoan lineage (Figure 2.1a). [30]

Briefly, the splicing process consists of two consecutive transesterification reactions (Figure 2.1b). In the first step, the 2′-hydroxyl group of a conserved adenosine residue found in the branch point sequence nucleophilic attacks the phosphate group of the guanosine in the 5′ splice site. This reaction results in the cleavage of the 5′ intron-exon boundary followed by the ligation of the terminal of the intron to the branch point in a lariat structure. In the second reaction, the 3′-hydroxyl group of the free 5′ exon (upstream exon) attacks the phosphate group of the 3′ splice site that creates the ligation of the adjacent exons and the excision of the intron. [30]

To compensate for the little information found in the intron sequence, splicing process occurs with the association of many *trans*-acting factors that will interact with the pre-mRNA to form the spliceosome machinery. In this way, the spliceosome will promote more efficient splicing by spatially positioning the reacting groups of the pre-mRNA for catalysis. Summarily, the spliceosome is gradually assembled and the splicing reaction is driven by RNA-RNA, RNA-protein and protein-protein interactions. The introns which the splicing process is driven by the spliceosome composed of the U1, U2, U5 and U4/U6 snRNPs and numerous non-snRNP proteins are said to be introns of the “U2-type”. Whereas, if the main-subunits of the spliceosome used for splicing were U11, U12, U5 and U4/U6 snRNPs, the intron is from the “U12-type”. [30] Basically, the differences found in both types of splicing process is that the U12-dependent splicing is characterized by a tightly constrained consensus sequence at the 5′ splice site, a longer and well-conserved branch point sequence and the absence of the polypyrimidine tract (Figure 2.2). [29,32] Initially, introns of the “U12-type” were named AT-AC introns due to the presence of AT and AC splice sites that appeared to break the rule of the invariant GT-AG splice sites of the “U2-type” introns. However, more extensive studies revealed that “U12-type” introns are mostly composed by the canonical GT-AG termini. In this way, the composition of the splice sites is no longer considered as a feature that distinguishes “U2-type” introns from the “U12-type” introns. [29,33]

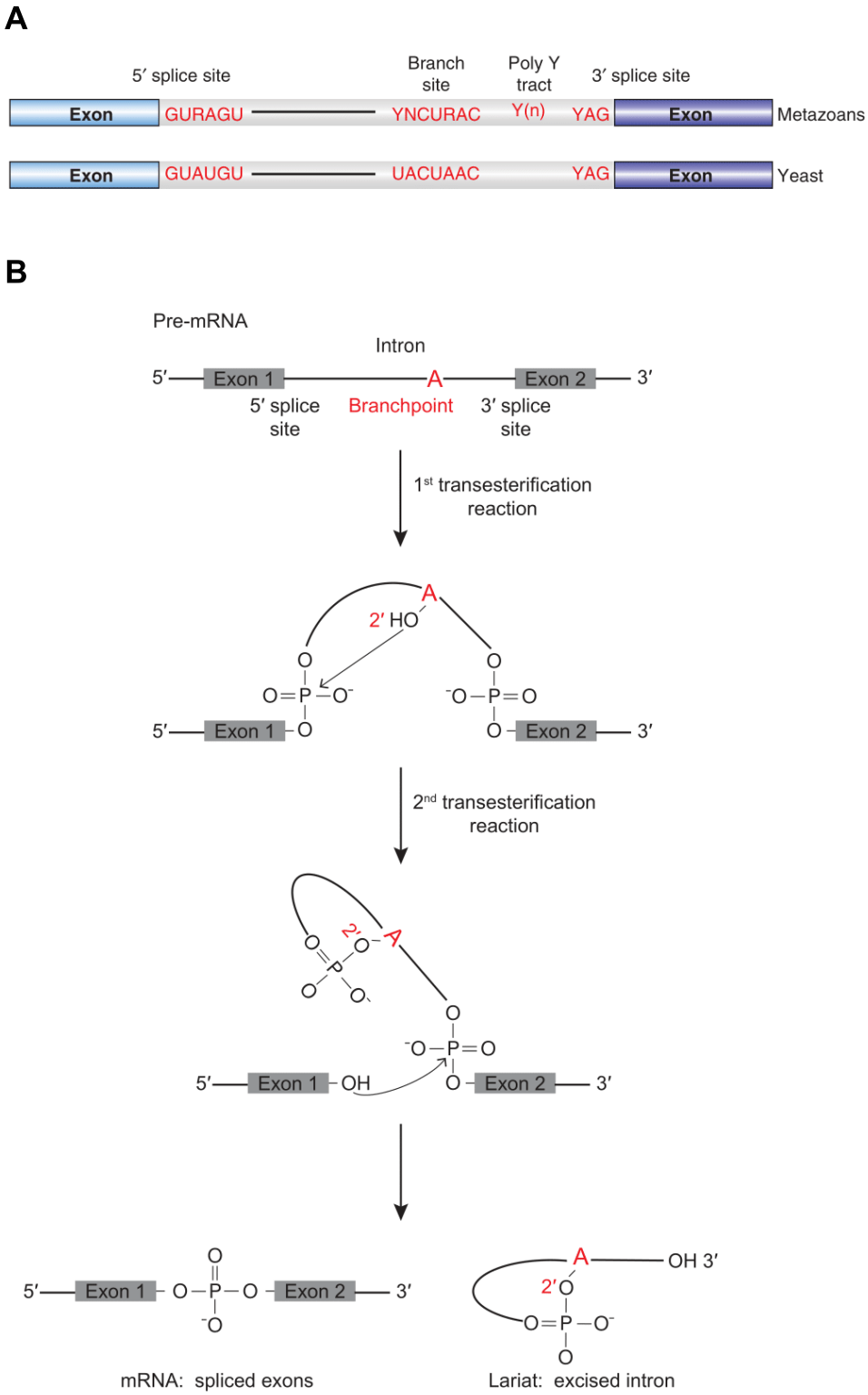


Figure 2.1 - Representation of the splicing process of spliceosomal introns. Commonly, introns are characterized by the presence of GU and AG dinucleotides in its 5′ and 3′ boundaries respectively, and a branch point sequence containing a defined A nucleotide (a). In metazoans, is also seen a polypyrimidine tract before the 3′ splice site (a). The splicing pathway can be simplified by two consecutive transesterification reactions (b). The first reaction consists of the break of the 5′ exon-intron boundary and the formation of a lariat structure in the intron sequence. The second reaction performs the ligation of exons and complete intron removal.

Source: (a) WILL; LÜHRMANN; [30] (b) SAINI. [31]

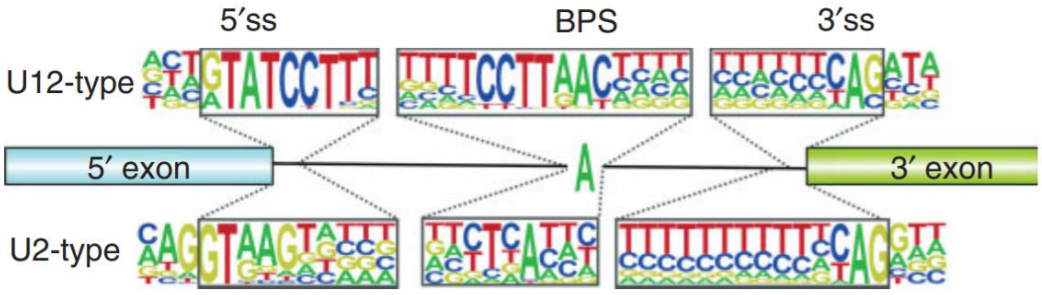


Figure 2.2 - Schematic representation of the splice sites in different types of spliceosomal introns. The spliceosomal introns can be classified as “U12-type” and “U2-type” based on the splicing machinery involved in their excision. The region near the splice site at the 5′-end of the “U12-type” introns, shown as “5′ss” in the top of the figure, is characterized by the presence of a tightly constrained consensus sequence. The branch point sequence of this group of introns (represented as the “BPS” box) is longer and well-conserved. However, this type of introns does not show a characteristic polypyrimidine tract at the end of the intron (“3′ss” box). Splice sites from “U2-type” introns, shown in the bottom, are known for its long and more conserved polypyrimidine tract, mainly composed of thymines, and also by its well-conserved GT and AG splice sites.
Source: TURUNEN *et al*. [32]

The “U2-type” introns are the most abundant type of introns found in the metazoan genes. In the U2-dependent splicing, the well-characterized U1, U2, U4 and U6 snRNPs are functionally analogous to the U11, U12, U4 and U6 snRNPs of the U12-dependent splicing, respectively. The splicing mechanism of the introns from the “U2 type” occurs in a stepwise manner (Figure 2.3). Firstly, the 5′ splice site is recognized by the U1 snRNP by an almost perfect Watson-Crick base pairing between the 5′ termini of the U1 small nuclear RNA (snRNA) and the consensus sequence of the exon-intron junction. Then, the branch point sequence is recognized by a non-snRNP factor known as SF1, and the polypyrimidine tract and 3′ splice site interact with the U2AF. Until this point, the assembly formed is known as the spliceosomal commitment or E complex. Later, the SF1 is displaced by the U2 snRNP that interacts with the branch point sequence using ATP to form the A complex, also known as the pre-spliceosome. The interaction of the U2 snRNP with the branch point sequence makes the adenosine residue exposed to later nucleophilic attack. The next step is marked by the recruitment of the complex composed of the U5 and the base-paired U4-U6, that together is known as the tri-snRNP complex, to form the spliceosome B complex. At this stage, the U4-U6 snRNPs unwind allowing the U6 to be base-paired with the U2 and the 5′ splice site, and the U5 interacts via base pairing with the 5′ and 3′ splice site. This process releases the U4 and U1 from the spliceosome assembly and forms the active spliceosome. Then, the spliceosomal complex catalyses the first

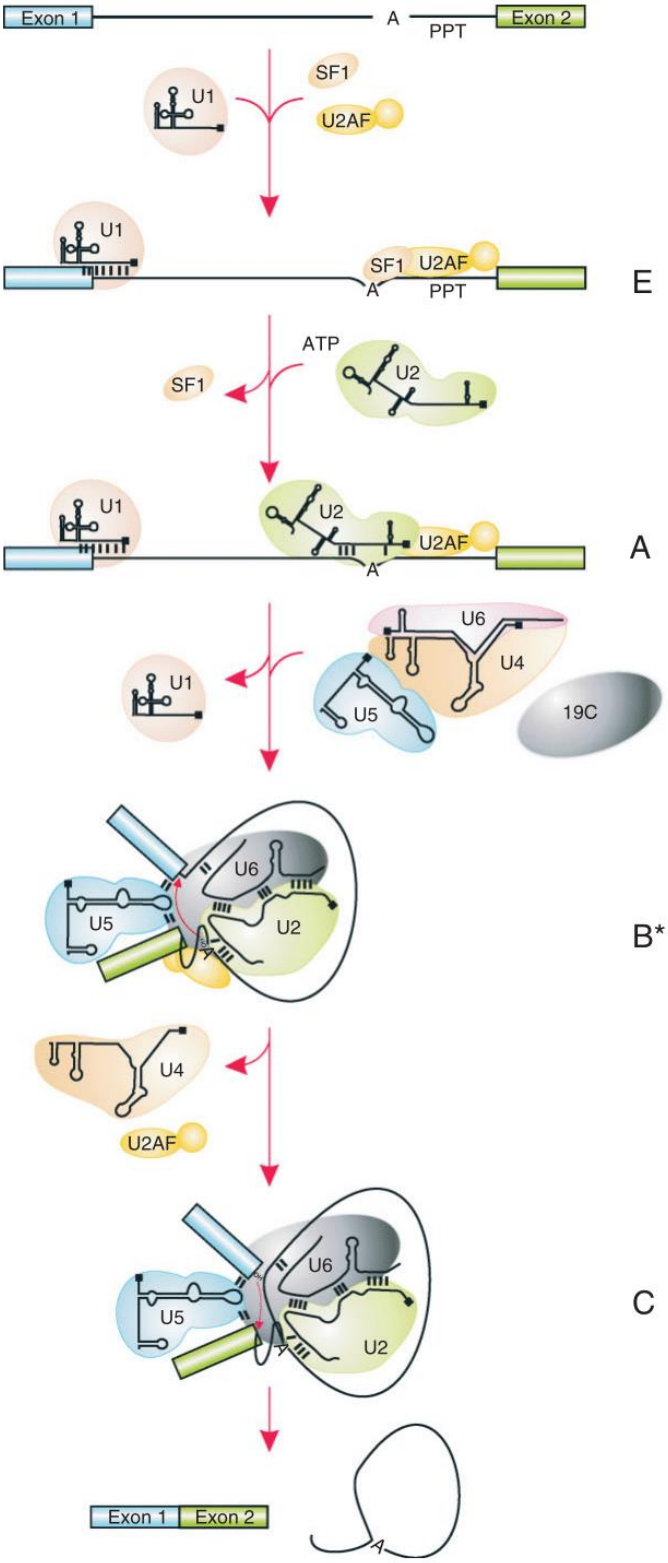


Figure 2.3 - Assembly of U2-dependent spliceosomes. The U2-dependent splicing is characterized by its stepwise nature. The spliceosome apparatus is gradually assembled to assist the intron splicing by spatially positioning the reacting groups of the pre-mRNA for catalysis. At the end of the two transesterification reactions, the intron is removed and the two adjacent exons are finally ligated. The schematic spliceosome proteins, snRNPs and their further interactions are represented. On the right, the name of the current complex is signed. The polypyrimidine tract is represented as “PPT”.

Source: TURUNEN *et al*. [32]

transesterification reaction, leading to the formation of a lariat structure in the intron. Subsequently, the second transesterification reaction occurs, probably with the help of the U5 by positioning the ends of the two exons. In the end, the intron is released in its lariat structure and the adjacent exons are ligated. [28–30,32]

The fashion of how the spliceosome is assembled depends on the two possible strategies used to identify exons from introns: if exons are first defined or introns are first defined. In lower eukaryotes, where introns are usually shorter than the adjacent exons, the spliceosome commonly identifies the intron first, whereas, in higher eukaryotes, where introns are usually much longer than the exons, the exons are usually first identified by the spliceosome machinery. Indeed, an important factor defining which assembly mechanism would be adopted is the length of the introns and exons. In this way, introns shorter than 250 bp are usually first identified in the splicing process, even if they are found in higher eukaryotes. Both modes of identification can be seen in the same pre-mRNA molecule. If the intron is first identified, the location of pairing between the splice sites takes place across the intron in a model that is called “intron definition”. On the other hand, if the exon is the one that is first identified, the pairing between the splice sites happens across the exon rather than the intron in a manner called “exon definition”. In the exon definition splicing, the communication between the spliceosome subunits occurs across the exon rather than the intron (Figure 2.4). However, it is important to keep in mind that both intron and exon definition modes of splicing do not differ mechanistically on a practical level. [28]

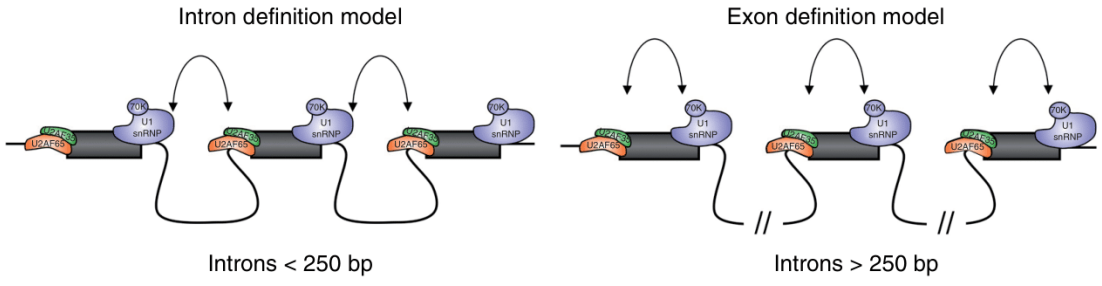


Figure 2.4 - The intron and exon definition models of splicing. The intron definition model occurs on introns shorter than 250 bp. In this model, the spliceosome machinery is assembled across the intron. On the other hand, introns longer than 250 bp are usually identified by the exon definition model where the splicing occurs across the exon. Both models can coexist in the same pre-mRNA molecule and their splicing mechanisms do not differ on a practical level. Introns are represented by a thin line and exons by a thick line. The arrows highlight the differences in the interactions between the spliceosome subunits of both definitions’ models.

Source: DE CONTI *et al*. [28]

2.2 The impact of spliceosomal introns in the eukaryotic genomes

Even though the spliceosomal introns compress one of the defined features of the eukaryotic lineage, its widespread proliferation remains a great enigma of the genomic architecture and evolution. [19,27] In humans, for example, a gene has 12.1 introns on average. [34] Even more surprisingly, more than 80% of human exons are shorter than 200 bp, [35] whereas the mean size of introns reaches more than 3,300 bp. [35-36] However, at the same time, the presence of introns in the genome is believed to impose potentially hazardous effects on the organism. There are three main described problems related to the nature of introns. First: the presence of introns is costly to the host cell, in terms of time and energy. In this way, the cell has to redirect energy to harbour and transcribe all the introns and the genes that code for the spliceosome - one of the largest molecular complexes in a cell with more than 150 proteins for the U2-type complex. Also, the presence of introns can considerably increase the necessary time for the transcription of a gene. On average, the RNA polymerase II elongation rate is 60 bases per second. [37] Therefore, for genes harbouring large introns, the transcription of a full-length mRNA molecule might last hours. [19] Second: changes in the splicing pattern might lead to genetic disorders, protein malfunctions or misproduction. In that context, defects at any step of the splicing process could result in a detrimental effect. [19] And third: the burden of critical introns recognition sites is related to the increased rate of null allele production through mutations. [27]

Taking into consideration all the deleterious effects described above, the prevalence and the great number of introns in the eukaryotic genome is astonishing. As questioned by Lynch: “How, then, did the eukaryotic genome arrive at the point at which ‘genes in pieces’ has become the norm?” [27] It is suggested that the simple fact of having a transcribed part of the genome that is free from a constricted selection can increase the genetic diversity that will probably result in the acquisition of many introns functions. [19,38] In fact, today, the roles of introns extend from the direct ones to even the indirect ones that makes them beneficial, or even essential, for their host cells. [19,39]

Interestingly, genomic portions corresponding to introns have important functions in the structure of the genome. For example, it is stated that introns are

related to the correct chromatin assembly, where introns function as depleted regions for the occupancy of the nucleosome. [40-41] In fact, genes with deleted introns show that the ability of nucleosomes formation is perturbed. [42-43] Moreover, genomic introns can also modulate the expression level of their harbour gene. Large-scale analyses have shown that intron-containing genes tend to be more expressed than genes without introns in yeast and also in mammals. [44-45] Indeed, an experiment showed that constructs with introns were 400 times more expressed than constructs without introns. [46] The transcription modulation is believed to occur thanks to some sequences, usually found in the 5′ introns, that function as regulatory elements of the main gene promoter. [19] Indirectly, genomic introns are also important for enhancing the efficiency of natural selection by increasing recombination rate. In this context, long introns are said to be capable of increasing the recombination rate of two sites in two neighbouring exons that leads to a greater opportunity of two favourable alleles at linked loci to be located together. [47]

Furthermore, the importance of introns is made clear when considering the different roles of splicing in cellular biology. Some splice sites are only recognized in certain tissues, times or conditions, leading to different forms of splicing. These alternative forms of splicing allow the production of multiple proteins from a single gene. An impressive example is the Dscam gene of the fruit fly *Drosophila melanogaster* that can potentially generate 38,000 isoforms, a number larger than the total genes of this species. [48] This possibility of producing a large variety of isoforms is not a simple passive consequence of introns' existence. In fact, introns possess *cis*-regulatory motifs that are capable of modulating the spliceosome assembly positively or negatively on nearby potential splice sites. [49-50] The splicing process can also be related to regulations on the transcription level. In the bovine papillomavirus, for example, the type 1 genes are known for only being transcribed in the late stage of the infection. This is achieved by repression caused by the interaction of U1 snRNA with elements that resembles the 5′ splice site. [51]

The splicing process is also important even after the occurrence of intron excision and exon ligation, because of the marks left in the processed RNA. These traces of the splicing reaction are found near the exon-exon junctions, the exon junction complex (EJC) that serves as a memory of where the introns used to be located. The

EJC is described for participating in many cellular processes related to the mRNA. For example, mRNA with a termination codon that resides 50-55 bp far from an EJC is targeted as a premature transcript. This transcript is later recognized and degraded by the nonsense-mediated decay (NMD). Originally, the NMD pathway was thought to be only a surveillance mechanism in eukaryotic cells that prevents translation of erroneous transcripts. However, recent studies have shown that the NMD is also a mechanism used in post-transcriptional regulation of gene expression. [52-53] The EJC is also related to facilitating the mRNA exportation from the nucleus to the cytoplasm by the recruitment of export factors. In fact, spliced transcripts are described as being exported faster than their unspliced counterparts. Overall, the exportation was enhanced 6-10-fold by splicing. [54] The presence of EJC in the processed mRNA has also been described for its importance in the recruitment of shuttle proteins that ensure the correct mRNA localization in the cytoplasm for translation. [19]

Surprisingly, even the increased transcription time has already been related to some of the introns functions. It was shown that an oscillatory transcription pattern can be resulted by a negative feedback with time delay. Interestingly, the obtained expression pulses have a cycle that depends on the length of the intron. [55] In mice, for example, the Hes7 gene responsible for somite segmentation has introns that cause a 19 minutes delay of transcription. Without those introns, the gene oscillatory transcription disappears and causes severe segmentation defects. [56]

2.3 Consequences of intron retention

Maybe one of the most evident advantages of the “genes in pieces" phenomenon is the ability to generate different forms of splicing, and thus a huge variety of transcripts that are capable of different functions. The alternative splicing was first described 40 years ago. At that time, it was discovered that a single gene could encode for a membrane-bound protein and also secreted antibodies. [57-58] In the beginning, the alternative splicing of transcripts was thought to be an unusual event. However, it is now acknowledged that alternative splicing is the norm in complex

eukaryotes. In humans, for example, nearly 95% of the multi-exonic genes are said to undergo different forms of splicing. [59]

The alternative splicing made possible the confection of many transcripts from a restricted number of genes. This resulted in the expansion of both transcriptome and proteome of eukaryotes. The alternative splicing can be achieved by many modes of splicing. The four basic types are: alternative 5′ splice site selection, alternative 3′ splice site selection, cassette-exon inclusion or skipping, and intron retention (Figure 2.5). [60] The mechanisms underlying the regulation of alternative splicing involves also *cis* and *trans*-acting factors, such as splicing enhancers or silencers, chromatin modifications, RNA secondary structure and more. [61–63] However, the details of how splicing decisions are determined have remained elusive.

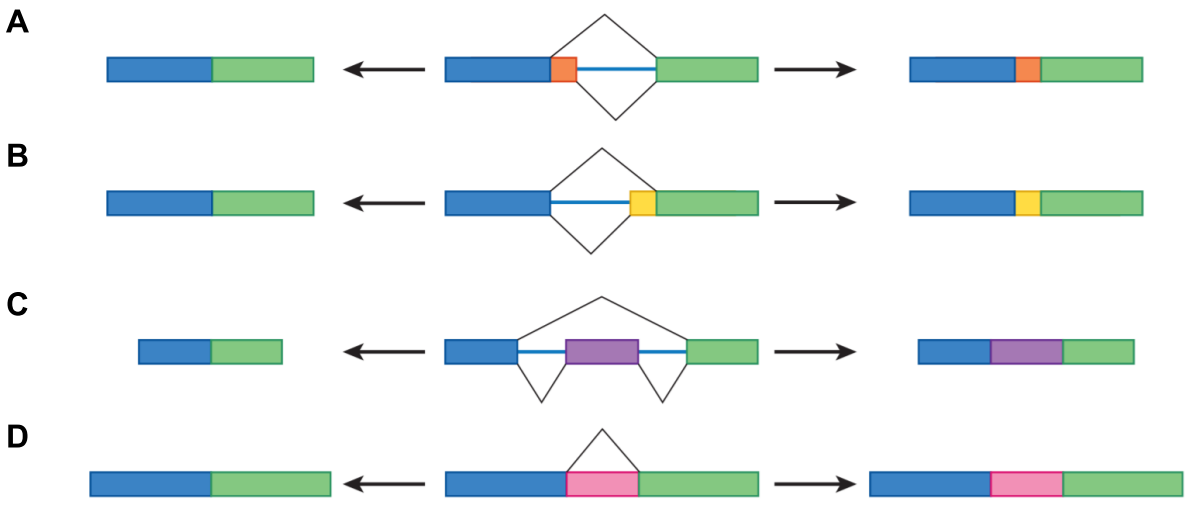


Figure 2.5 - The four types of alternative splicing. The different forms of creating alternative splicing by alternative 5′ splice site (a), alternative 3′ splice site (b), exon inclusion or skipping (c), and intron retention (d) are shown. In the middle, the representations of the forms of splicing are highlighted by the back lines. On both sides, right and left, there are different results of mRNA depending on the form of the previous splicing. If the represented central portion of the mRNA is removed, the mRNA will correspond to the mRNA shown on the left side. However, if the central portion is included, the resulting mRNA will be like the ones shown on the right.
Source: NILSEN; GRAVELEY. [60]

There is still a lot to learn about the mechanisms behind the alternative splicing. The intron retention (IR) is maybe one of the less understood modes of alternative splicing creation. For long, it was considered to be derived from errors in the splicing process, being mostly ignored in studies with mammals. [64-65] However, recent studies

have shown that IR plays an important regulatory role in protein isoforms production, RNA stability and efficiency of the translation process. Furthermore, the fact that transcripts with a retained intron can be stored at the nucleus and afterwards be quickly processed might represent a mechanism for prompt induction of expression. Also, the presence of one or more retained introns in the processed transcript can redirect this molecule to distinct pathways, providing another layer of regulation to the host cell. The presence of transcripts with introns in the cytoplasm may trigger several regulatory processes. For example: IR in the 5′ untranslated region (UTR) can activate or repress the initiation of translation; IR in the 3′ UTR may promote a stabilization effect in the mRNA; introduction of premature termination codon (PTC) due to IR can lead to the production of truncated proteins or transcript degradation via NMD; introduction of additional codons by IR can result in the production of new protein isoforms. Furthermore, transcripts with IR can be retained in the nucleus for posterior degradation by the exosome pathway or even to be safely stored until external *stimuli* induce late splicing (Figure 2.6). Interestingly, IR tends to be more frequently seen within short introns, where the splicing occurs via the “intron definition” model. On the contrary, long introns governed by the “exon definition” model tend to promote exon skipping in case of incorrect processing. [66-67] In a study involving the introns of 23 species of fungi and protists, it was shown that the correlation of IR and intron length is evident. [68] The same was also seen in mammals. [69–71]

Nowadays, the importance of IR in some complex biological processes is being explored using sophisticated methods of analysis. In spermatogenesis, for example, it has been shown that IR is the most common form of alternative splicing. In this context, it is hypothesized that such high production of IR transcripts has the objective of safely storing transcripts in polyribosomes until posterior expression days after their synthesis, due to the inability of these cells to actively produce transcripts from their condensed chromosomes during meiosis. [72] In addition, IR has also been described for its importance to circumvent situations of cellular stress. In plants, IR is already well known as one of the main mechanisms used for regulating the expression of genes and for its use in abiotic stresses situations. In *Arabidopsis thaliana*, for example, IR induces the formation of isoforms with PTC in response to high salinity environments. [73-74] More recently, studies in mammals have also demonstrated the use of IR in situations of cellular stress. During heat shock, increased IR events were observed in

about 1,700 genes whose transcripts were retained in the nucleus. This population of genes were enriched for tRNA synthetase, nuclear pores and spliceosomal functions. However, a set of 580 genes enriched for proteins involved in oxidation-reduction pathways and protein folding showed co-transcriptional splicing and no IR. [71]

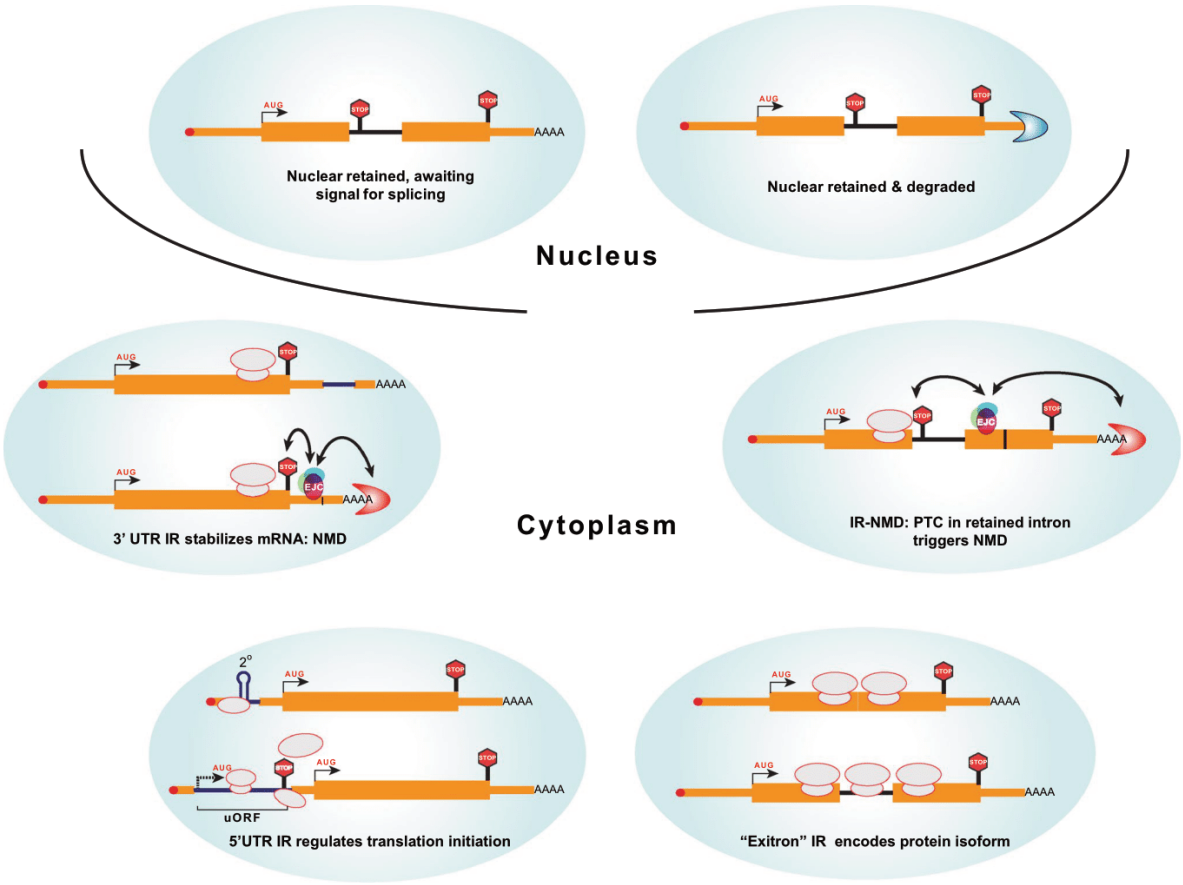


Figure 2.6 - The consequences of IR on an mRNA molecule. Cytoplasmic and nuclear transcripts do show different fates. In the nucleus, the transcript could be stored waiting for splicing signals, or degraded. In the cytoplasm, retained introns can stabilize the transcript, regulate the translation initiation, lead the transcript to degradation via NMD, or even be translated into a protein isoform. Introns are represented by a black line, exons as thick orange blocks, UTR as orange lines, stop codons as the "stop" flags, and the 5′ cap as the red circle.
Source: JACOB; SMITH. [65]

Due to its involvement in many intricate biological processes, the dysregulation of IR events is being described as correlated to some pathological conditions. A study involving 16 different types of cancer showed that IR is the form of alternative splicing that differ most between normal and cancerous tissues. An increase in IR was reported to be related to most of the studied forms of cancer, except for breast cancer that shows a reduction in the number of IR cases. [75] In fact, data suggest that an important

issue involving IR is its dysregulation in pathogenic states, not simply its rise or fall in number. In a study of the relationship of IR with the ageing process and the advent of Alzheimer's disease in mice, a dynamic process of rising and fall of IR cases in the brain tissues according to the age of the studied mice was observed. Changes of the IR pattern during ageing was reported to regulate the transition from healthy to a pathological condition in late-onset sporadic Alzheimer's disease. [76]

Chapter 3

THE MINIMAL INTRONS OF EUKARYOTES

One of the striking features of spliceosomal introns is the apparent lack of conserved features in most of their sequences, with loose conservation of few short sequence motifs. Such lack of conservation is also reflected in the size of spliceosomal introns, which range from ten bases to millions of bases. [77] At first sight, such a broad length range might suggest a very loose constraint in introns sizes. However, analysis of the distribution of introns sizes in diverse organisms showed an accumulation of introns near the minimum size observed for the organism. The introns belonging to this population were termed “minimal introns” and their existence suggests the presence of evolutionary forces acting to limit sizes in a subset of introns. The reason why this particular population of introns is subjected to such evolutionary forces and what their functions are is still under study. In this chapter, the definition, prevalence and functions of this group of introns will be examined.

3.1 A look into the introns size

Introns are poorly understood gene entities. For long considered as “junk DNA”, introns were left behind due to its lack of conservation seen on their degenerated sequences or even their length. In eukaryotes, one of the smallest introns with only 30 bp long were identified in the human MST1L gene that codes for the putative macrophage stimulating 1-like protein. [78] On the other hand, the DhDhc7(Y) gene, which protein is essential for the motility of the sperm tails and fertility of the *Drosophila hydei* fly male, harbours an impressive long intron greater than 3.6 Mbp. [79] This massive difference in intron length reveals there are not strong forces acting to constraint intron sizes at a particular length, or they are acting differently in subsets of

introns. In cases of very short introns, a particular constraint is the necessity of some elementary sequences that make possible their recognition, which explain a threshold for minimum intron size observed in many eukaryotic genomes. The very large size of some introns may be explained in part by the presence of transposable elements or satellite DNA. However, given the evident energetic burden in maintaining and transcribing such large introns, it is not clear why they are not selected against.

Even though multiple sizes of introns are observed, there is an apparent favoured intron length in the studied eukaryotic genomes. When analysing the intron size distribution of multiple species from different kingdoms, it was revealed a frequent accumulation of introns near the minimum size. Interestingly, the existence of a mode in introns' length extends from many evolutionary stems of the eukaryotic lineage, but its size is specific for each species (Figure 3.1). [34,77] A favoured intron length is observed in representatives of protozoa (Figure 3.1a), fungi (Figure 3.1b), metazoan (Figure 3.1c), plants (Figure 3.1d) and vertebrates (Figure 3.1e-f). In plants and vertebrates, the accumulation of introns in a minimum size is represented by a well-conserved intron size parameter (Figure 3.1d-f). The second peak in vertebrates represents an intron expansion phenomenon, where there is an increase in the proportion of larger introns in the genome (Figure 3.2e-f). The second peak is a consequence of the representation of the distribution of introns sizes on a logarithmic scale.

Those introns found under and near the characteristic peak of the species are commonly mentioned as minimal introns. Its name refers to the apparent impossibility of having introns smaller. It is hypothesised that the minimum size of introns reflects the physical constraints imposed by the splicing machinery and the need for some critical sequences in order to still be recognized for splicing. In fact, the observation that minimal introns peak change according to the species may be explained by the variations in the spliceosomal machinery and their dimensions on different species. [34] Furthermore, it was observed that minimal introns are under constraints of selective pressure for maintenance of their optimal sizes. In humans, for example, it was shown that introns longer than the peak size of 87bp have a higher ratio of deletion over insertion than those introns that are shorter than the optimal size. [80]

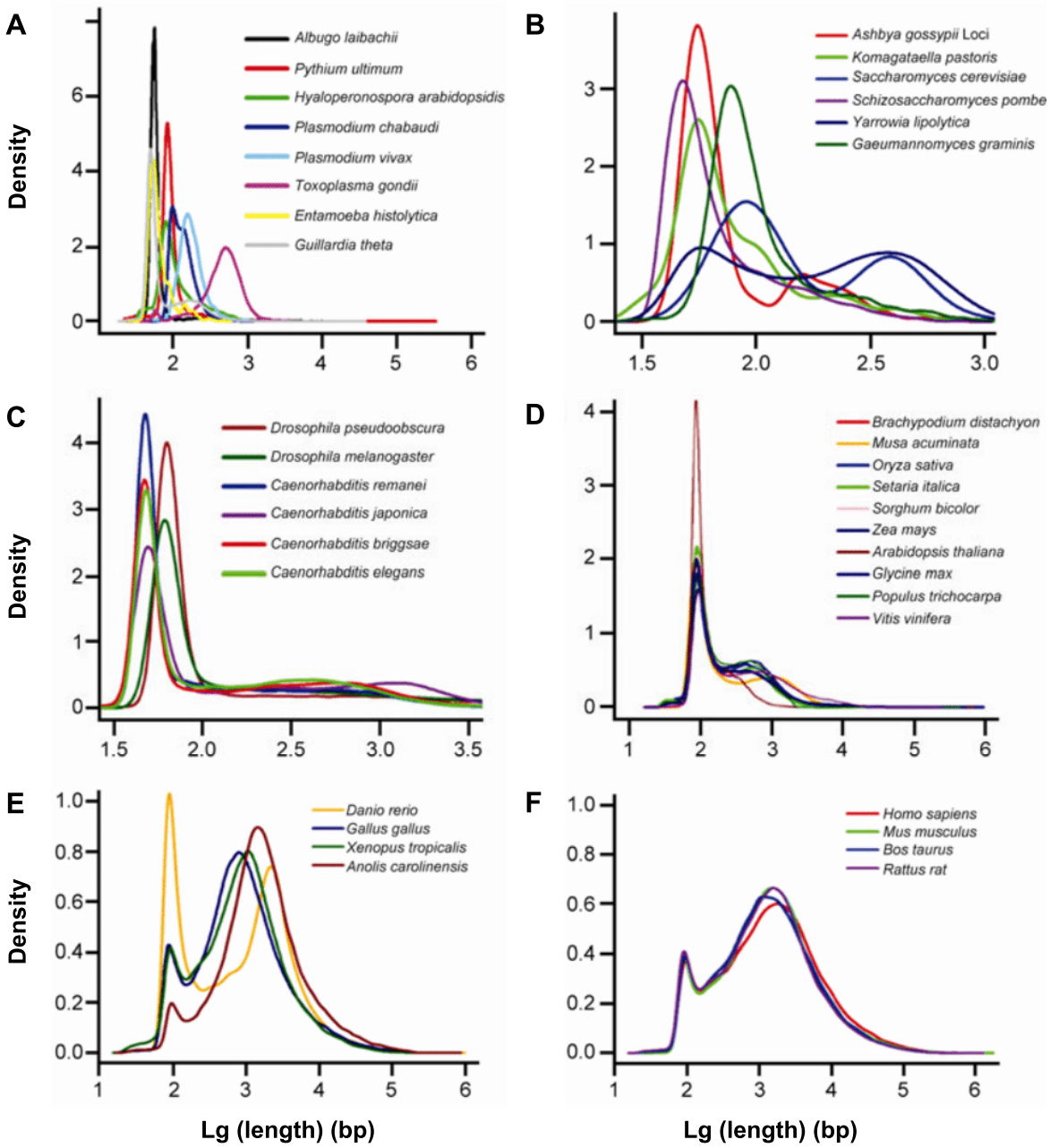


Figure 3.1 - Introns size distribution of diverse eukaryotic groups. Introns were analysed from representatives of protozoa (a), fungi (b), metazoa (c), plants (d), vertebrates without mammals (e), and mammals (f). The introns size modes are represented by a peak which is specific for each species. Lengths are shown on a logarithmic scale of base 10.

Source: JIAYAN *et al.* [77]

3.2 The features of minimal introns

The well-shared existence of minimal introns in many eukaryotic species, together with the reported selective pressure for maintenance of their optimal size,

raised the question: is there a functional reason for minimal introns persistence? In this context, the other intriguing question that might be asked is: do minimal intron-containing genes share unique features that differentiate them from the other genes? In the following paragraphs, it will be presented some evidence that corroborates for the affirmative answer of both questions.

Based on the known prevalence of the neutral theory of evolution, [81] the first hypothesis to be elaborated would be that there is no functional need for the existence of minimal introns. Following this argument, maybe the observed intron peak would be created by the fact that some introns were never bombarded by transposons, which is a known cause for expanding introns sizes, or even if they were, those introns have deleted back to their sizes as mutations biased on deletions is higher than insertions. Together with the fact that introns might show a minimal size given the splicing machinery constraints, the peak would be explained without any functional reason. So, if this hypothesis is correct, we shall see that the process of creating minimal introns would not be favoured for any intron or gene. However, using the minimal introns found in the *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans* genomes, scientists have shown that the presence of minimal introns is biased towards a particular cluster of genes. This suggests the existence of two distinct populations of introns, of which only one of them is subject to evolutionary pressure for maintenance of introns at a small size. [34]

Other evidence points to that selection for small intron size in some genes may be associated with the functions observed in this subset of genes. Minimal introns have been related to the synthesis, splicing, and export of the minimal intron-containing genes. This type of gene is considered unique because of functional and genomic features. [77] Minimal intron-containing genes were found to be longer in size, localized preferentially in the vicinity of open chromatin and to be mainly related to housekeeping functions. [82-83] Interestingly, these genes are also said to replicate earlier and to be more efficiently exported from the nucleus to the cytoplasm. In this context, the "Routing Hypothesis" was proposed, which assumes that the minimal introns are considered as tags so their containing genes could be processed and routed differently from the other types of genes. [77] In this way, the highly abundant and large housekeeping genes that are located in the surface of the chromatin territories are

selectively exported. This might reduce the risk of being entangled with other genes located at the interior of the chromatin (Figure 3.2). [82]

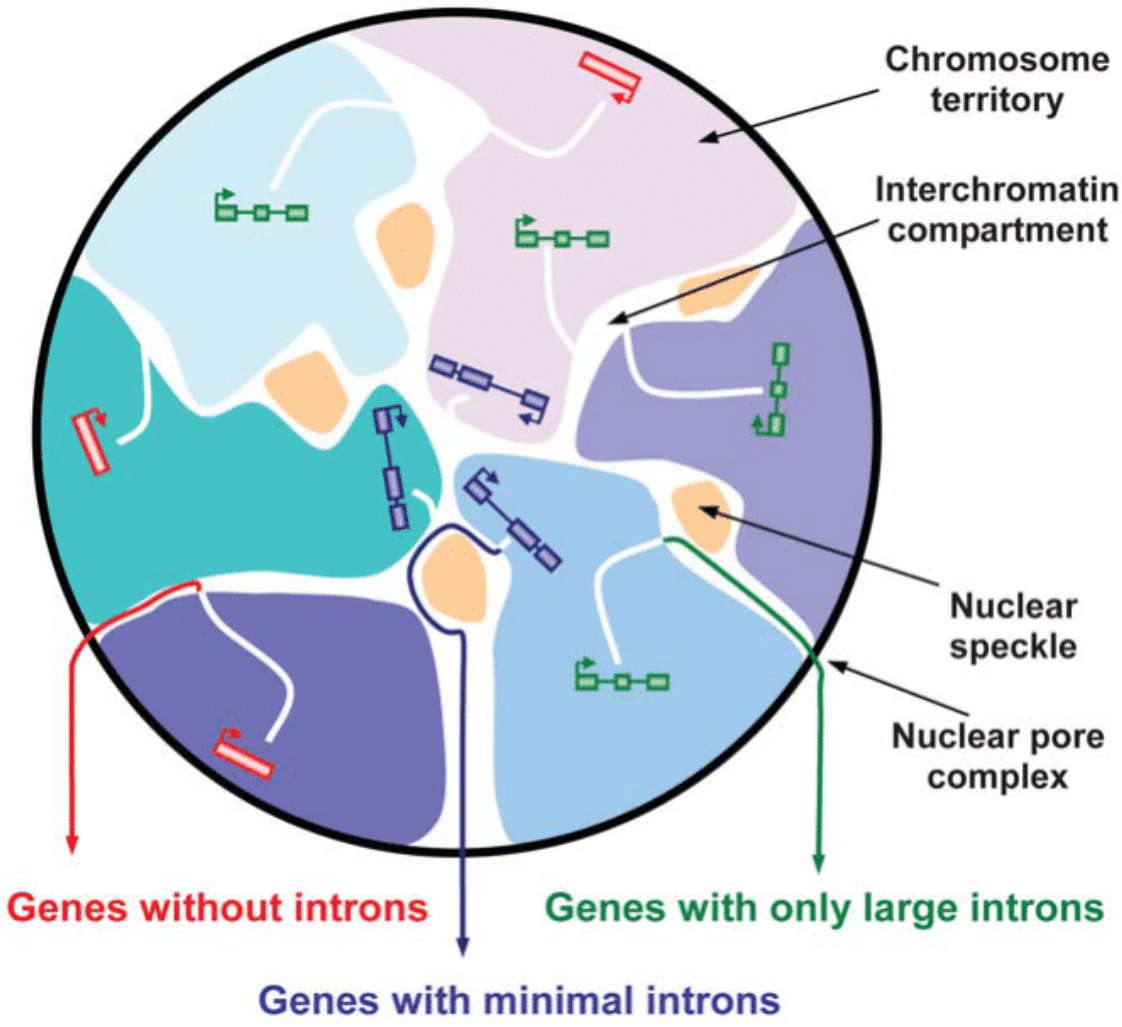


Figure 3.2 - Schematic representation of a mammalian nucleus showing the nucleocytoplasmic export route model. Minimal intron-containing genes are shown in blue. These genes are strategically localized on the surface of the chromosome territory for facilitating the splicing of their large and abundant transcripts. Transcripts derived from other types of genes show different export routes. The independent pathways of different genes prevent the entanglement of mRNPs and make the export process more efficient.
Source: ZHU *et al*. [82]

Part II

MINIMAL INTRONS OF THE PLATYHELMINTH *SCHISTOSOMA MANSONI*

The phylum Platyhelminthes, which name came from a Greek origin meaning flatworms, is a group of bilaterian animals dorsoventrally flattened with no body cavity other than a gut. With a low estimative of 30,000 species, the Platyhelminthes is among the most abundant invertebrates in the biosphere. [84-85] In this context, the Digenea group from the Trematoda class comprises approximately 18,000 nominal species [86] and it is by far the most diverse group of this phylum. [84,87] The Digenea subclass is, without a doubt, a successful group of internal metazoan parasites. [86] The majority of its members are endoparasites that affect all classes of vertebrates. [84]

The biomedical importance of the Digeneas is well exemplified by members of the *Schistosoma* spp*.* These species are the causative agents of the schistosomiasis, a neglected disease popularly known as bilharzia. Estimates show that around 290.8 million people need preventive treatment of schistosomiasis in 2018, mostly living in poor communities in Africa where access to safe drinking water and adequate sanitation is limited. The *Schistosoma* spp. parasites have a complex life cycle involving humans as the definitive hosts and freshwater snails as the intermediate hosts. In humans, infective parasite eggs are produced sexually, while in snails, the cercariae are produced via asexual route. The disease is characterized by the symptoms caused by the human body's reaction to the parasite eggs that include diarrhoea, lethargy, abdominal pain and the presence of blood in the faeces. [88-89] The schistosomiasis disease is responsible for 280,000 deaths annually, and a worldwide burden of 3.3 million disability-adjusted life years. [90]

In this section, we aim to study the minimal introns from the Platyhelminthes phylum. To conduct our study, we decided to choose the *Schistosoma mansoni* as a model organism, mainly because this species has a well-studied genome due to its medical importance as the primary and widespread causative agent of the schistosomiasis disease. [91] Our analysis is divided into three chapters: "Distribution of

minimal introns in the genome of *Schistosoma mansoni*", "Sequence characterization of minimal introns from *Schistosoma mansoni*" and "Intron retention of minimal introns from *Schistosoma mansoni*". Each chapter begins with a brief introductory section that discusses some important concepts needed for the understanding of the conducted experiments, followed by "Results and Discussion" and "Materials and Methods" topics. At the end of this section, the "Conclusion" chapter will present the most pertinent ideas concerning the minimal introns of the chosen genome.

Chapter 4

DISTRIBUTION OF MINIMAL INTRONS IN THE GENOME OF *SCHISTOSOMA MANSONI*

In eukaryotes, diverse cellular processes such as gene recombination, transcription and translation might represent a good source for adaptation. In this context, the dynamics of those processes might be reflected in the architecture of chromosomes and genes. It has been shown, for example, that low recombination rates of genes located in centromeres regions of the chromosome might promote faster divergence and facilitate the speciation process. [92] Sequences located in different parts of the gene itself are also under distinctive selective pressures as particular steps of transcription and translation occur in distinct parts of the gene and the mRNA. In fact, it is shown that the position of intron might influence its function on regulating transcription initiation, enhancement of splicing efficiency, chromatin assembly, the efficiency of mRNA export and recognition of premature termination codons via NMD pathway. [93]

Knowing the influence of genomic entities positions on the evolution of the eukaryotes, our study of minimal introns from the *Schistosoma mansoni* parasite begins with a brief statistical description in relation to its number, prevalence in the genome, gene and chromosome locations. The results of this first section will give us a panorama on the distribution of minimal introns that will work as a background for making further speculations on the role of those introns in the genome.

4.1 Minimal introns of *Schistosoma mansoni* are remarkably short

The analysis of intron length from the *Schistosoma mansoni* resulted in a typical

distribution, with a minimal peak similar to that observed in many other species of eukaryotes (Figure 3.1 and Figure 4.1). However, the modal length of introns in this species is at 34 bp, that configures a relatively low number when compared to what is commonly observed in other eukaryotic species (Figure 3.1, Table 4.1). For the purpose of our analysis, minimal introns in this species were defined as those with a length between 31 bp to 43 bp (see “Intron size distribution” topic in “Materials and methods”). Introns longer than 43 bp were considered as regular-sized introns.

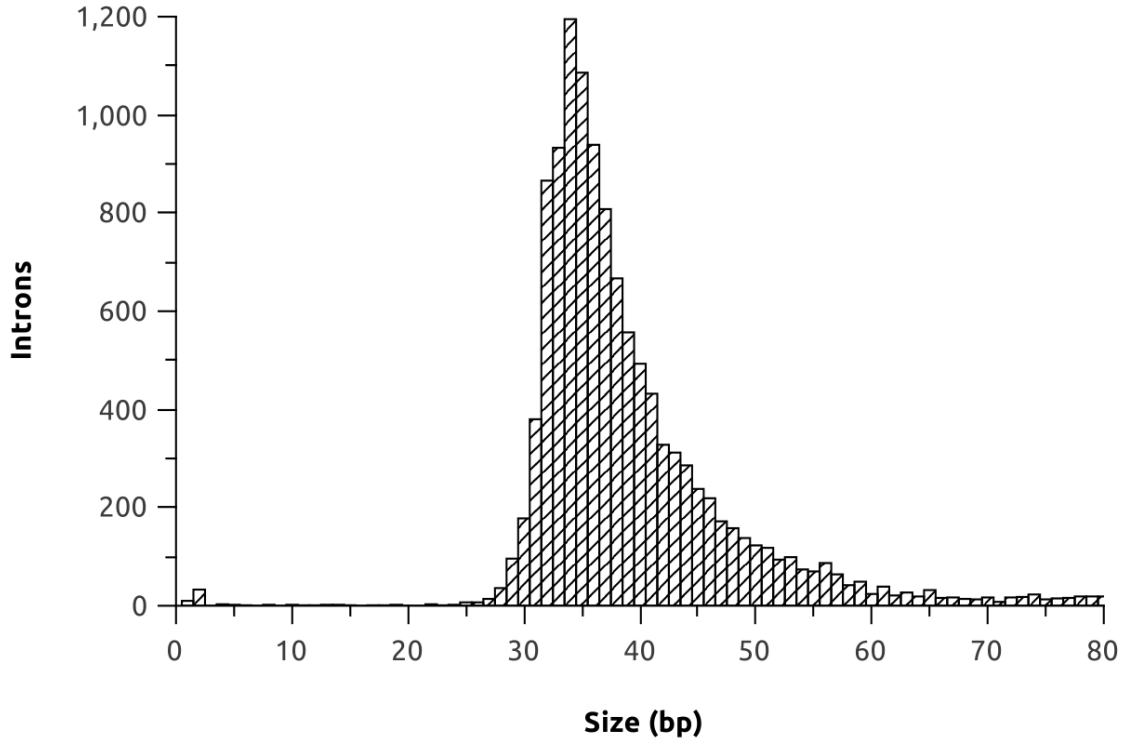


Figure 4.1 - Distribution of minimal introns’ length from the *Schistosoma mansoni* genome. As observed in multiple eukaryotic species, there is a frequent accumulation of introns near the minimum size. The modal length is at 34 bp and the minimal introns were defined as those from 31-43 bp. Introns longer than 43 bp are referred to as regular introns. The peak shows a sharp decline on lengths shorter than the modal length. The x-axis was truncated on 60 bp length, however, it extends up to 182,204 bp.
Source: By the author.

Interestingly, the minimal introns peak does not fit into a normal curve distribution. Instead, the peak is characterized by a sharp decline in the frequency of introns shorter than the modal length. We suggest that this observation might be derived from a negative selective pressure present on introns shorter than a minimum size needed for the correct identification of the intron. This selective pressure might be

the result of a physical constraint that renders very short introns unrecognizable to the splicing apparatus. Therefore, we would expect that this pressure would be stronger than the one responsible for the deletions in introns longer than the modal length that is also needed for the maintenance of the introns' peak.

Table 4.1 - Modal length of minimal introns of eukaryotes.

Species	Minimal Intron Modal Length (bp)
Homo sapiens [a]	92
Arabidopsis thaliana [a]	89
Drosophila melanogaster [a]	61
Caenorhabditis elegans [a]	48
Schistosoma mansoni [b]	34

Source: (a) YU *et al*.; [34] (b) By the author.

4.2 Minimal intron-containing genes comprises half of nuclear genes

Analysis of the *Schistosoma mansoni* genome shows that 74,298 introns are distributed in 10,117 nuclear genes, with an average of 7.3 introns per gene. It was previously reported that this number is 12.1, 6.2, 4.7 and 7.7 for the *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans* genomes respectively (Table 4.2). [34]

We verified that 8,996 introns, or 12.1% from the total number of nuclear introns of the *Schistosoma mansoni genome*, display lengths between 31 bp to 43 bp and are distributed in 4,082 genes. On this way, from the 9,476 intron-containing genes - around 93.7% of total nuclear genes, nearly 43.1% have minimal introns with a mean of 2.2 minimal introns per gene.

Table 4.2 - Introns per gene from *Schistosoma mansoni* and previously analysed species.

Species	Introns per Gene
Homo sapiens [a]	12.1
Caenorhabditis elegans [a]	7.7
Schistosoma mansoni [b]	7.3
Arabidopsis thaliana [a]	6.2
Drosophila melanogaster [a]	4.7

Source: (a) YU *et al*.; [34] (b) By the author.

Table 4.3 - Prevalence of minimal introns and minimal intron-containing genes in the genome of *Schistosoma mansoni*, two plants and six vertebrates species.

Species	Minimal Introns (% of total introns)	Minimal Intron-Containing Genes (% of total genes)
Homo sapiens [a]	10.6	32.1
Mus musculus [a]	10.4	25.6
Gallus gallus [a]	12.6	44.8
Anolis carolinensis [a]	5.6	30.5
Xenopus tropicalis [a]	13.8	51.6
Danio rerio [a]	25.7	60.3
Schistosoma mansoni [b]	17.6	51.3
Arabidopsis thaliana [a]	72.3	57.1
Oryza sativa [a]	47.2	51.8

Source: (a) JIAYAN *et al*.; [77] (b) By the author.

In order to compare our results with other animals and plant species from a previous study, the above analysis was also conducted on minimal introns defined as introns shorter than 150 bp. A total of 13,072 minimal introns were detected on 5,195 genes. When analysing the fraction of minimal introns and minimal intron-containing genes in the *Schistosoma mansoni* genome with two plants and six vertebrates species, our data shows more similarity with the results observed in vertebrates representatives (Table 4.3). As suggested by JiaYan *et al*., [77] larger introns have been proportionally increasing through evolutionary time scales in the vertebrates' history. Interestingly, this might also be the case of the *Schistosoma mansoni* genome when compared to plant genomes. Whereas plants show a high fraction of minimal introns in the genome, our Platyhelminth representative shows an analogous result to those observed in vertebrates.

4.3 Minimal introns are preferentially found on the 5′ end of genes

As previously discussed in the introductory section, sequences located in different parts of the gene are under distinctive selective pressures. It was already shown that this observation can be well exemplified by the influence of the intron position on its function. [93] In order to understand how intron position might influence the introns from *Schistosoma mansoni*, we decided to analyse the intron size distribution on each intron position and compare our results with previously reported data.

As suggested by Berriman *et al*., [94] the genome of *Schistosoma mansoni* differs from the other eukaryotes by the presence of a skewed size distribution along genes where introns closer to the 5′ end are smaller than the introns at 3′ end (Figure 4.2). However, when applying the same methodology to our introns dataset, our result showed two main differences. The first difference is that the median size of introns, instead of its mean values, appears to be skewed towards small sizes near the 5′ end of the gene (Figure 4.3a). Introns mean size values distribution in relation to the 5′ end appears to be more convoluted. We observed a large average size of introns in the first position, followed by a substantial decrease, and afterwards, values appear to be

somewhat stable (Figure 4.3b).

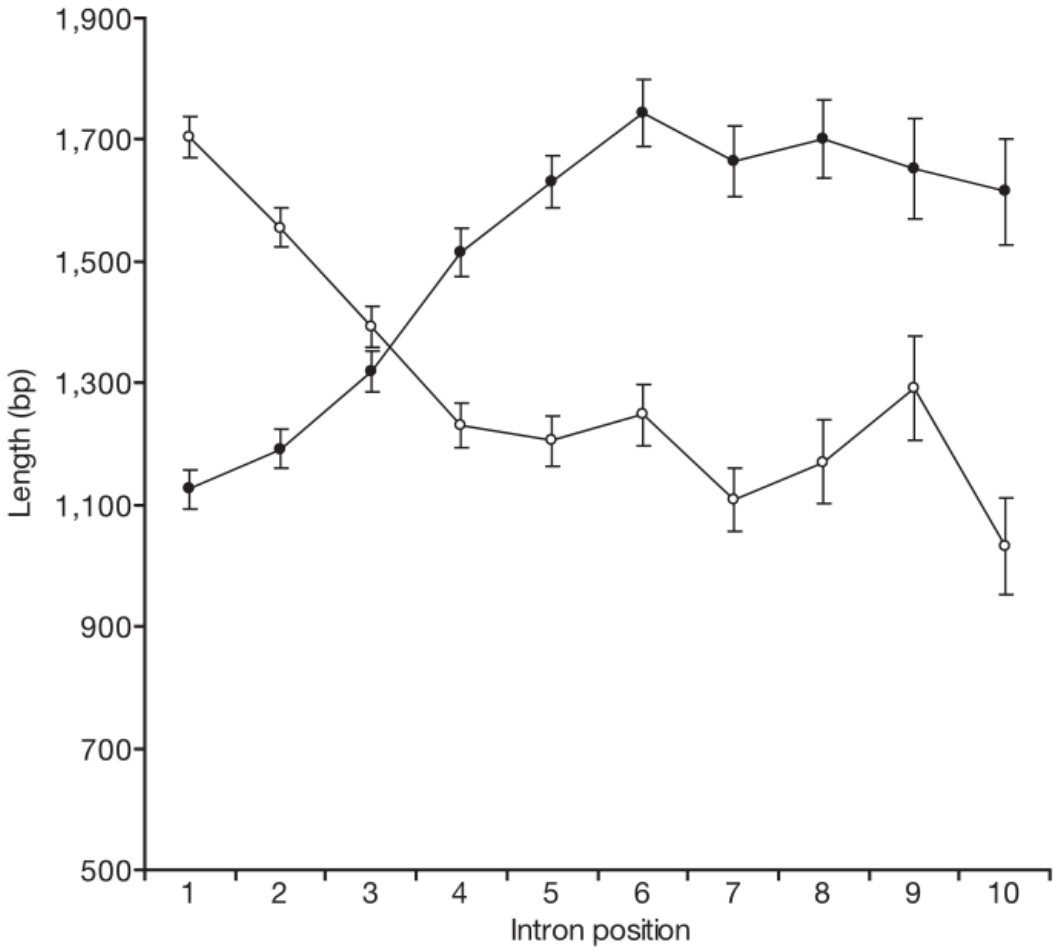


Figure 4.2 - Previously reported intron size distribution along the transcript. *Schistosoma mansoni* introns show a skewed size distribution according to intron positions. At the 5′ end, introns tend to be smaller. Mean lengths ± standard errors are shown. Positions relative to the 5′ end and 3′ end are represented by the solid and open circles, respectively.
Source: BERRIMAN *et al*. [94]

It is also interesting to observe that, for all positions, mean values tend to be much higher than median values. This suggests the presence of outliers of very high length that tend to have a strong influence in the mean but not median values. Therefore, the sharp peak at the first position near the 5′ end in the mean intron size that is not paralleled by the median, the second difference found in our result, is suggestive that a larger portion of first introns presents exceptionally large sizes or the size of these outliers is larger than in other positions. One possibility to explain such a considerable size difference is that first introns tend to be located in UTRs. As observed in *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Arabidopsis thaliana*,

the media intron size is greater at the 5′ UTR than at the coding sequence (CDS) region of the gene. [93]

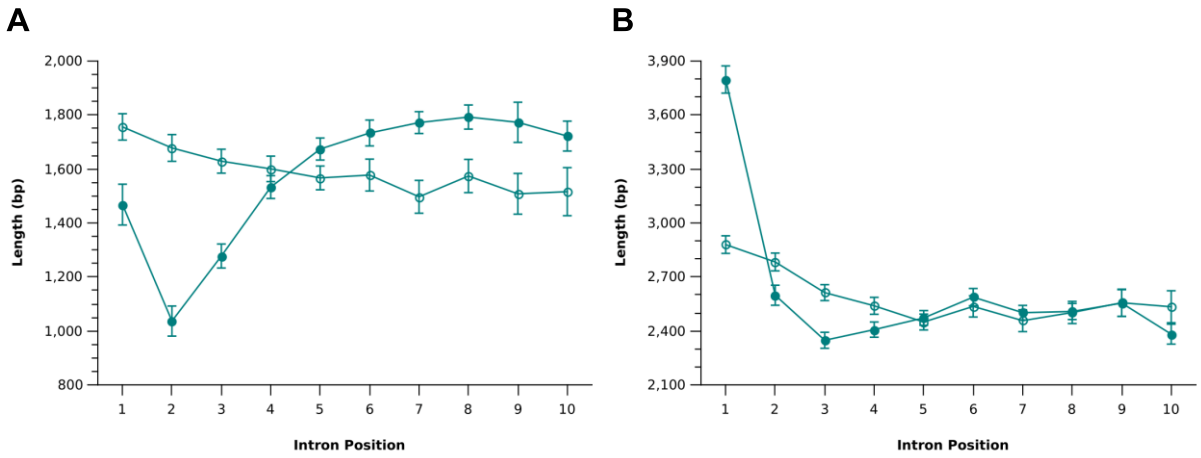


Figure 4.3 - Intron size distribution along the genes. Intron length varies in accordance with its position in the gene. Positions were counted relative to the 5′ end and 3′ end and are represented by solid and open circles curves, respectively. Median (a) and mean values (b) of intron sizes are shown together with standard errors. Except for the first position relative to the 5′ end, median intron sizes tend to be smaller (a). On the other hand, intron size distribution according to intron positions is not well represented by mean measures (b).

Source: By the author.

In order to better explore the observed skewed intron size distribution along genes and the effect of 5′ UTR on this analysis, we decided to compute the fraction of minimal introns along all intron positions and CDS positions only. We suggest that the lower median size values obtained towards the 5′ end of the gene might be affected by a greater presence of minimal introns at those positions. When analysing the first 10 possible intron positions in the gene, UTR included, we observed that the second position from the 5′ end, which was previously known for a sharp size decrease, shows the largest fraction of minimal introns in the whole gene (Figure 4.4a). This intron position harbours 2,606 minimal introns and counts for approximately 30% of all introns found in this position. On the other hand, the first intron position from the 3′ end has 504 minimal introns that represent around 15% of all introns located at this position. Interestingly, when removing the 5′ and 3′ UTRs from the analysis, we also observed the removal of the peak found in the previous second intron position (Figure 4.4b). This corroborates to the reported data of larger intron sizes detected at the 5’ UTR in relation to the CDS region of genes.

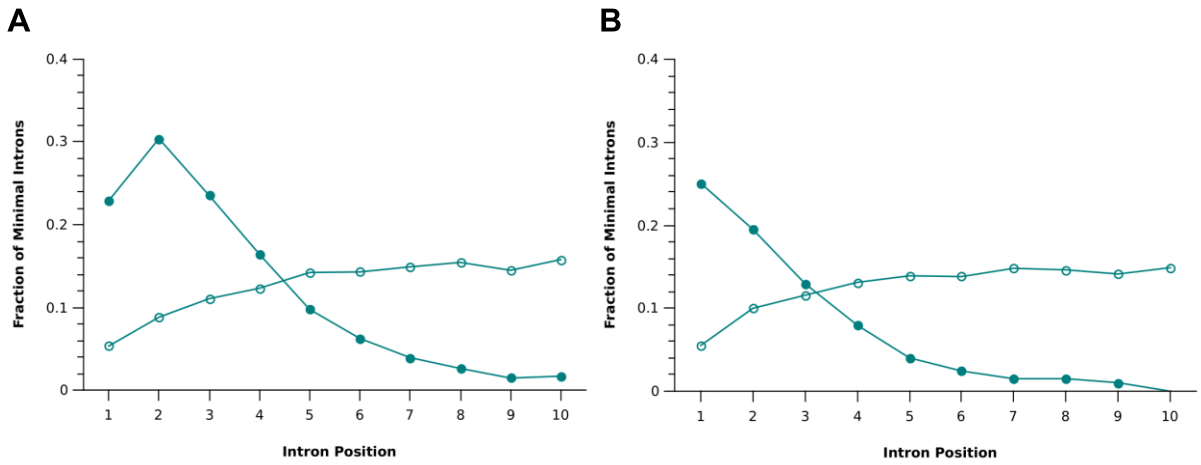


Figure 4.4 - Fraction of minimal introns in relation to the total number of introns along the genes. Intron positions were determined relative to 5′ end and 3′ end as shown by the solid and open circle curves, respectively. All parts of the gene (CDS and UTR) were considered in the analysis shown in (a), whereas only the CDS region was computed in (b). Minimal introns tend to accumulate at the 5′ end of the gene. However, the 5′ UTR might be populated by a smaller fraction of minimal introns as the peak on the second position shown in (a) was not observed when analysing only the CDS part of the gene (b).
Source: By the author.

Table 4.4 - Fraction of minimal introns in the chromosomes of *Schistosoma mansoni*.

Chromosome [a]	**Minimal Introns (%)**
1	12.3
2	13.4
3	14.0
4	13.2
5	7.8
6	9.4
7	6.1
Z/W	12.5

[a] Data from contigs are not shown due to its short length.
Source: By the author.

4.4 Minimal introns are enriched in some chromosomes

In our previous analysis, we observed that minimal introns comprise nearly 18% of all annotated introns in the genome of the blood fluke *Schistosoma mansoni*. We now aim to investigate if there is a differential distribution of minimal introns on chromosomes or chromosome regions of the genome of *Schistosoma mansoni*. For this, we calculated the fraction of minimal introns in and across the annotated chromosomes/contigs. Minimal introns showed a prevalence for 1, 2, 3, 4 and Z/W chromosomes with a mean percentage of approximately 13% of all introns found on those locations (Table 4.4). However, using a sliding window of 5 Mbp with a step of 1 Mbp, no preferences towards the central or extremes of the chromosomes were detected (Figure 4.5).

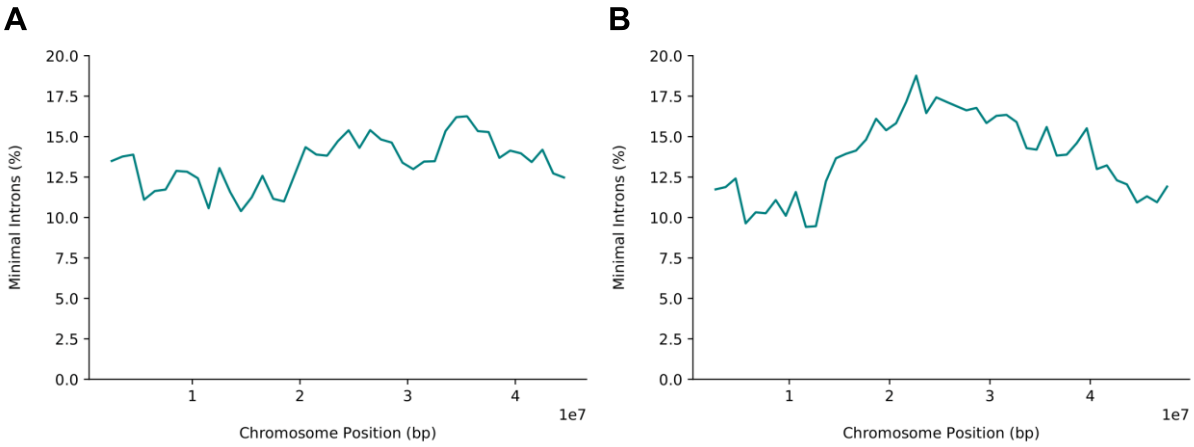


Figure 4.5 - Fraction of minimal introns along the chromosomes from *Schistosoma mansoni*. Using a sliding window of 5 Mbp and a step of 1 Mbp, the percentage of minimal introns from the total genes of each chromosome/contig were calculated. Examples of minimal intron fraction across chromosome 2 and 3 are represented in (a) and (b) respectively. No preference towards central or external regions of the chromosome was detected.
Source: By the author.

4.5 Materials and methods

4.5.1 Intron size distribution

For this and other analysis, we used the most recent genome data from the

Schistosoma mansoni, version 7, available as FTP link at the Sanger Institute database: ftp://ftp.sanger.ac.uk/pub/project/pathogens/Schistosoma/mansoni/v7. Data of coordinates of the genomic elements were extracted from the genome annotation file (GFF extension) via customized scripts. Introns positions were deduced based on the described exons coordinates. The selection of introns in case of multiple transcripts from a single gene was performed following the criteria: introns were selected from transcripts with the highest number of exons and the greatest transcript length, on this order of priority. Later, the size of introns was calculated given their coordinates. After establishing the distribution of introns based on their lengths, the range of the minimal introns' peak was defined as those introns' size near the minimal observed length of the genome with at least a population of 25% in number of the population from the modal length. The low limit definition step aims to remove very small introns that may represent annotation artefacts of the genome. Were considered as regular-sized introns, those introns whose size was greater than the upper limit of the minimal introns' range. For this and other tasks, Bash language (version 4.3.48) and Awk text processing library (version 4.1.3) were used unless stated. Graphs were created using QtiPlot software (version 0.9.8.9).

4.5.2 Fraction of minimal introns and minimal intron-containing genes in the genome

After extracting the information of introns length and location from the genome annotation file, the total number of introns, minimal introns, genes and minimal intron-containing genes were obtained. In order to compare our results with previous studies, besides using the criteria of minimal introns as 31 bp to 43 bp, we considered to our analysis the classification of introns shorter than 150 bp as minimal introns as well.

4.5.3 Introns size distribution along genes

In order to study how introns are distributed along the genes, distribution of intron sizes according to their position in the gene were analysed. The number and fraction of minimal introns per position were computed. The position of introns in relation to the 5′ end and 3′ end of the respective gene was computed. The first ten positions were used in the analysis, even if the gene harbours more intron positions.

The placement of the intron in a UTR or CDS was used as a classifier to distinguish possible distinct evolutionary pressures acting according to the context of the intron.

4.5.4 Minimal intron distribution along the chromosomes

For each chromosome or contigs found in the *Schistosoma mansoni* annotation file, the number of total and minimal introns was computed for the calculation of the minimal intron's frequency relative to the total number of introns. A sliding window approach was also used to calculate the relative frequency throughout each chromosome. To this end, the chromosome was scanned using a 5,000,000 bp sliding window with a 1,000,000 bp step. The relative frequencies were used instead of absolute intron numbers to avoid the influence of differential gene density. The analysis was conducted using Python programming language (version 3.7.3) with the support of Pandas (version 0.23.4) and Matplotlib (version 2.2.3) libraries for data manipulation and plotting visualization improvement.

Chapter 5

SEQUENCE CHARACTERIZATION OF MINIMAL INTRONS FROM *SCHISTOSOMA MANSONI*

The splicing process can be summarized as the mechanism in which an intron is removed and the adjacent exons are ligated in the RNA molecule. In eukaryotes, this process is catalysed by the spliceosome, a complex made of proteins and snRNPs that detects and orchestrates the removal of introns. There are four known recognition marks inside introns that are used for its detection by the spliceosome: a 5′ and 3′ splice sites found in the border of the intron - commonly characterized by the presence of the GU and AG dinucleotides respectively; a branch point sequence found near 18-40 nucleotides upstream the 3′ splice site; and also a polypyrimidine tract between the branch point sequence and the 3′ splice site that is found in the metazoan lineage. [30]

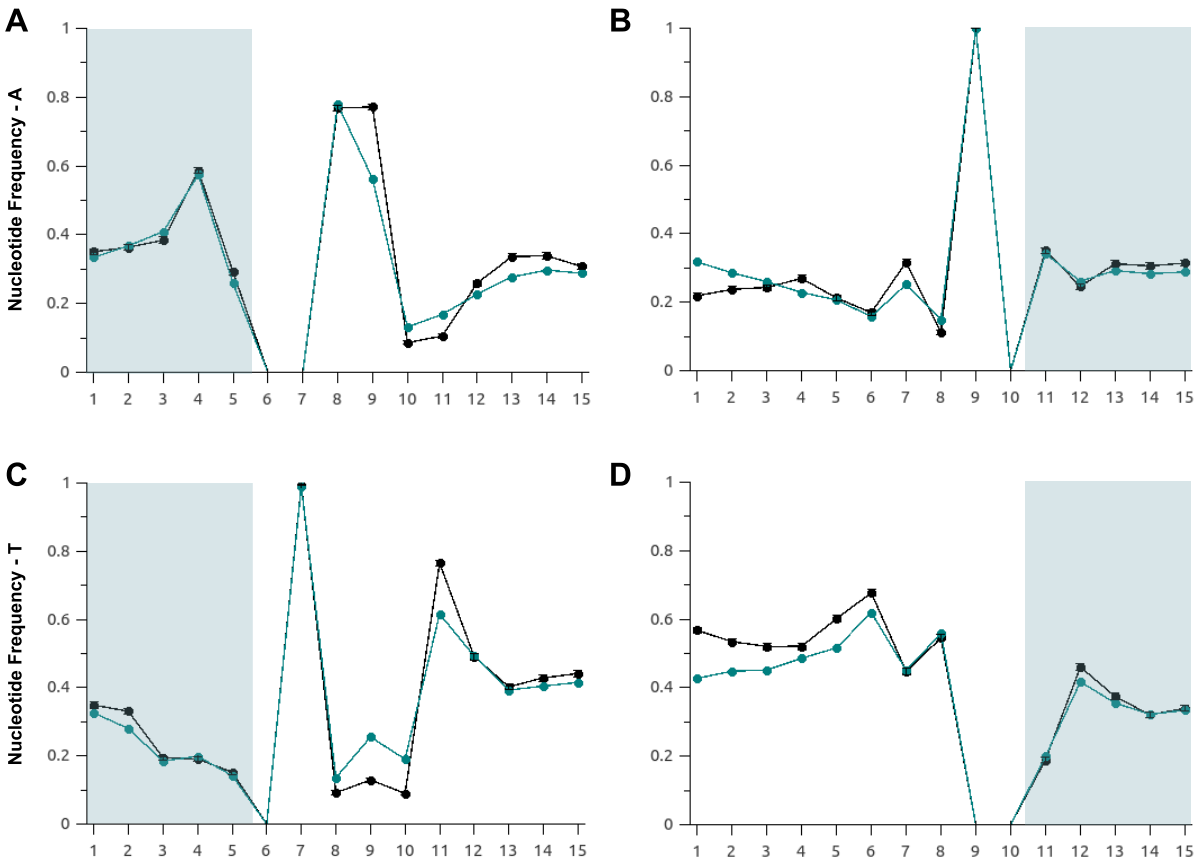
In this chapter, we investigate the nucleotide proportion and sequences motifs found in the minimal introns from *Schistosoma mansoni* that may be important components for the splicing process. Differences and similarities between the results of the analysis of minimal introns and regular-sized introns from the *Schistosoma mansoni* will be discussed. Our results will also be compared to the observations found in introns of other species whenever possible. We hope this study may contribute to the understanding of how such small minimal introns are recognized by the spliceosome, highlighting any particular feature this type of introns may present.

5.1 Minimal introns show typical U1- and U2-type splice sites

From the pool of recognition marks needed for the correct splicing of introns, the splice sites are the most conserved in terms of location and sequence. [95] In this

way, our investigation on the sequences of minimal introns from *Schistosoma mansoni* begins with the identification of their splice sites. This analysis also permits us to verify the conservation of the intron sites annotated in the *Schistosoma mansoni* genome.

In this analysis, introns were classified according to its length as minimal (intron length between 31 bp to 43 bp) or regular (if greater than 43 bp). The frequency of all four nucleotides in each position of the boundaries of exons and introns was calculated for both groups of introns (Figure 5.1). On the last two intronic positions of the 3′ end of minimal introns, A and G nucleotides predominate (frequency of approximately 1 for both nucleotides). Interestingly, the AG acceptor splice site is commonly preceded by T (frequency of 0.56), or less frequently by C (0.29) or A (0.15). Similar frequencies are observed in regular-sized introns. In humans, however, the AG is commonly preceded by C (0.65 cases of U2-type splice sites), or T (0.29 cases of U2-type splice sites), and rarely by A (0.06 cases of U2-type splice sites). [96]



(continued)

(continuation)

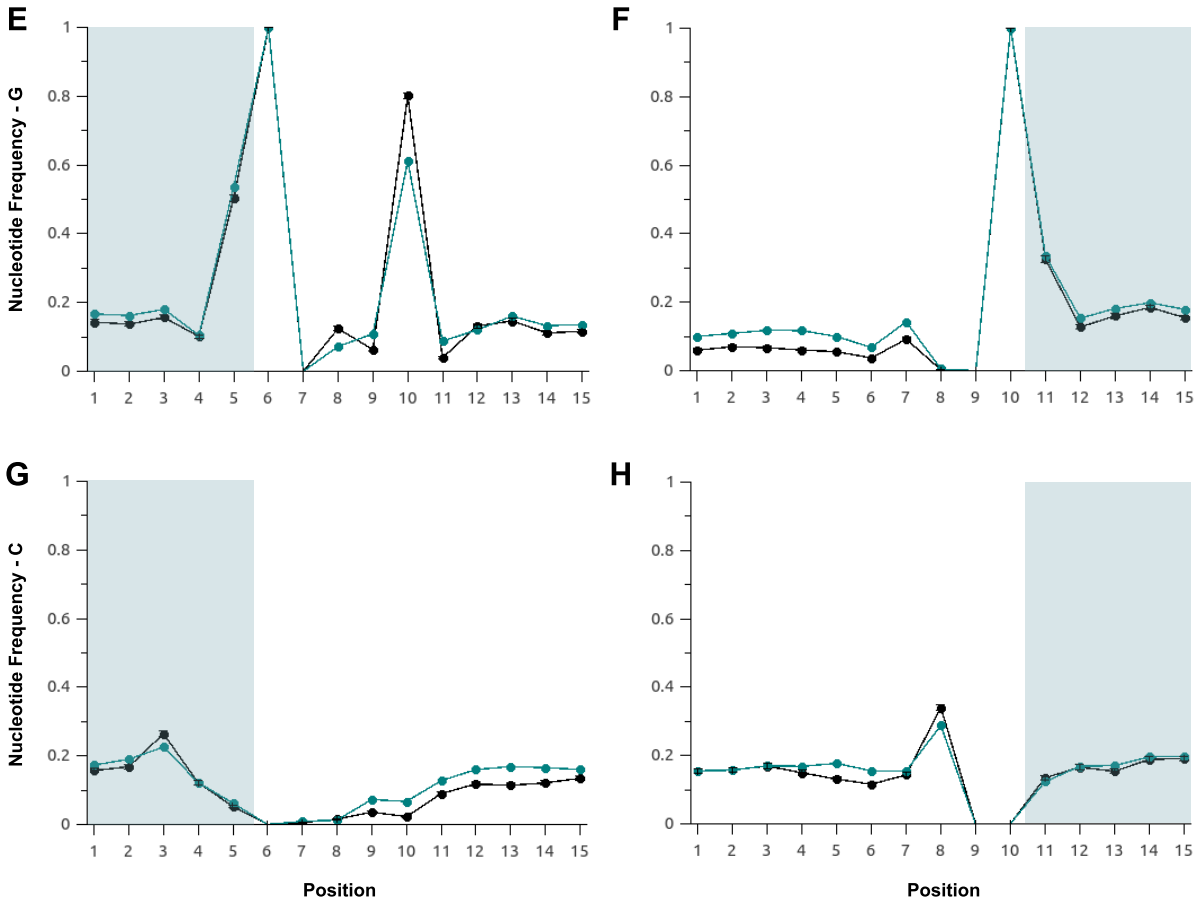


Figure 5.1 - Nucleotide frequency of the border regions between minimal introns and adjacent exons. 5′ and 3′ ends of minimal introns from *Schistosoma mansoni* are shown in the left (a, c, e, g) and in the right side (b, d, f, h) charts, respectively. Adenosine frequency is shown in (a, b), thymidine in (c, d), guanosine (e, f) and cytidine in (g, h). The nucleotide is referenced in the y-axis title. Intronic positions are represented by a transparent background and exonic positions by a coloured background. Data from minimal introns are shown in green and regular-sized introns in black. Bars represent the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) derived from 10,000 simulations where an identical number of regular-sized introns were sampled. At the 5′ end of both regular-sized and minimal introns, the splice site is characterized by the high frequency of the dinucleotide GT in the first two intronic positions, followed by AAG nucleotides and preceded by exonic AG nucleotides. Another GT is detected at the fifth and sixth intronic positions. At the 3′ end, the nucleotides AG found in the last two intronic positions represents the acceptor splice site. This AG is commonly preceded by the T nucleotide.
Source: By the author.

In the first and second intronic positions at the 5′ end, the G nucleotide is followed by the T nucleotide in almost all minimal introns (frequency of approximately 1 for both nucleotides). The same donor 5′ splice site was detected in regular-sized introns. Similar to what is observed in humans, GT dinucleotides are commonly followed by AAG and preceded by AG nucleotides found at the 3′ end of the upstream

exon. Mutations on the G found in the fifth intronic position have been related to reduce the splicing efficiency in yeast. [97] This consequence is believed to be a result of its importance in the initial recognition of the 5′ splice site by U1 and after spliceosome rearrangement, by U6. [98-99] Curiously, however, the fifth G is commonly followed by T in the sixth intronic position. If this second GT represents an alternative splice site, the reading frame of the transcript will be probably changed as the distance between the first and second GT is not a 3 multiple and no other possible splice site (AG) was found in the 3′ end. This could result in the translation of an erroneous transcript.

In order to measure the strengths of the splicing sites of both groups of introns, the quantity of information was calculated for each position of the exon-intron boundaries. This metric is expressed by bits, where 0 bits is equivalent to an equal proportion of nucleotides, and 2 bits is the highest quantity of information where all sequences have the same nucleotide at the analysed position. GT and AG splice sites found in the 5′ and 3′ extremities of the introns, respectively, are indicated to be highly conserved sequences in minimal and regular-sized introns (Figure 5.2). On the other hand, the second GT of the fifth and sixth intronic positions at 5′ end shows a lower sequence information. Curiously, this number is lower in minimal introns than regular-sized introns. In fact, this tendency of less conserved sequences of minimal introns predominates through all other analysed positions.

5.2 Minimal introns of *Schistosoma mansoni* lack a clear polypyrimidine tract

The polypyrimidine tract is one of the main *cis*-sequences needed for intron splicing. It is normally found between the branch point and 3′ splice site. This recognition mark interacts with U2AF and it is important for the spliceosome assembly. [30] Different sequences, positions and length of polypyrimidine tract could affect the efficiency of branch point utilization, selection of alternative branch sites, and 3′ splice site recognition. [100] Mammalian species have a clear polypyrimidine tract, however, the same is not observed in yeast. [100] In this section, we aim to identify the sequence and location of the polypyrimidine tract of minimal introns from *Schistosoma mansoni*.

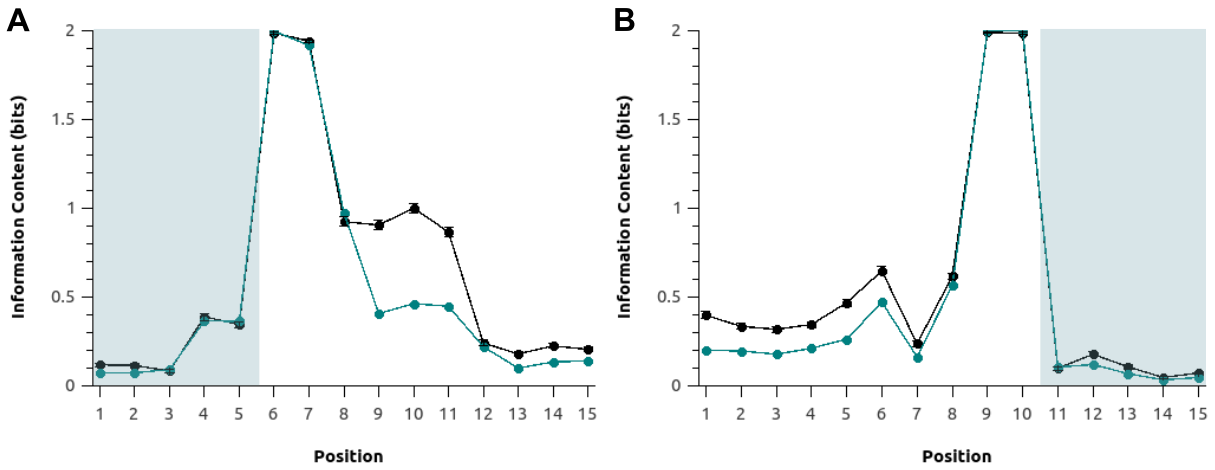


Figure 5.2 - Information content of the border regions between introns and adjacent exons. 5′ and 3′ ends of introns from *Schistosoma mansoni* are shown in (a) and (b). Intronic positions are represented by a transparent background and exonic positions by a coloured background. Data from minimal introns are shown in green and regular-sized introns in black. Bars represent the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) derived from 10,000 samplings where a number of regular-sized introns identical from that of the minimal introns were sampled from the total pool of regular-sized introns. The information content of minimal introns is lower when compared to regular-sized introns. High information content is seen in the first two (5′ end) and the last two (3′ end) intronic positions. This content is generated by the bases GT and AG respectively, which characterize the splice sites.
Source: By the author.

The polypyrimidine tract is reported to be mainly composed by uridines instead of cytidines. [101] Indeed, when analysing the nucleotide proportion of different minimal introns, we observed a preference for uridines (represented by T in the genomic DNA) throughout the intron. However, this preference is even higher in the immediate vicinity of the 3′ splice site (Figure 5.3a). In order to check if this increase of uridines is relevant, the quantity of information for each position of intronic sequences plus five positions of the adjacent exons was calculated. To remove the possible influence of the intron composition towards AT from this analysis, we also calculated a threshold of the quantity of information in a stochastic scenario where only the base content was maintained. To that end, we performed 10,000 simulations in which we shuffled each intronic sequence and computed the base composition of each position to obtain a set with the exactly the same base composition, but without any preference for a determined base in any position. We observed higher information content than that of the threshold in positions adjacent to the 3′ splice site that matches the previous observation of increased T proportion (Figure 5.3b).

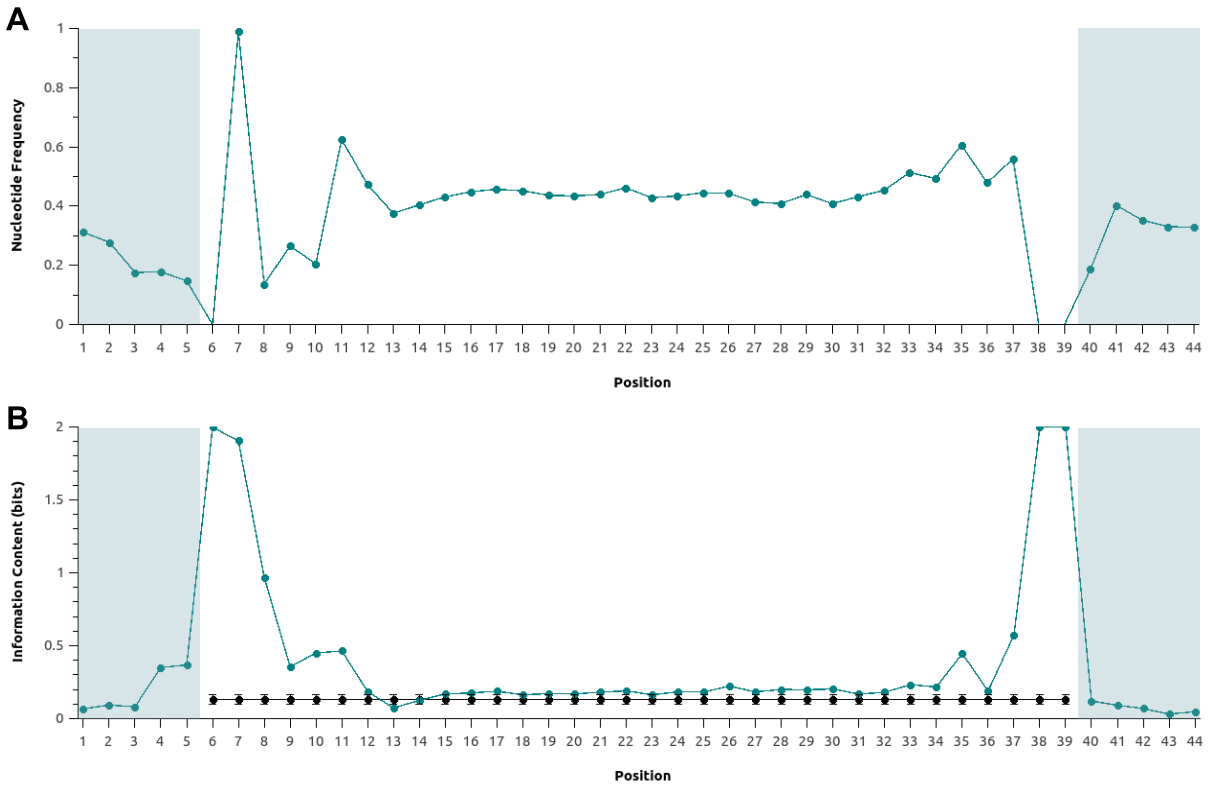


Figure 5.3 - Analysis of thymidine nucleotide proportion and information content within minimal intron sequences. Exemplified by 34 bp introns (length of minimal intron peak, n=1,195), the thymidine proportion (a) and the information content (b) was analysed along all intronic positions plus five positions of the adjacent exons (coloured in green). (a) The elevated thymidine proportion among all introns positions may reflect the high AT content of the *Schistosoma mansoni* genome. However, this proportion increases in the positions that precede the 3′ splice site that could indicate the presence of a polypyrimidine tract. (b) In fact, there is an information increase at positions adjacent to the 3′ splice site that matches the previous observation of increased thymidine proportion (green curve). The threshold curve in black (b) represents the information content obtained from the random shuffling of the intronic sequence. Bars represent the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) derived from 10,000 simulations.

Source: By the author.

As the polypyrimidine tract is described by a $(Y)_n$ consensus sequence, we decided to analyse if the observed high thymidine proportion in the intron sequence is caused by the presence of concatenated thymidines. To verify the number of uninterrupted thymidines and remove the probable influence of high AT content of introns, the AT skew of different subsequences of *k*-mer length was analysed (1<=*k*<=6). If greater than 0, the AT skew indicates that more A than T is found in the respective position. If lower than 0, the opposite is true. Values range from -1 to 1. Exemplified by the 34 bp introns (length of the minimal intron peak), the presence of concatenation of the same type of nucleotide is higher for T than A. In fact, the greater the *k*-mer analysed, the greater the tendency for T (Figure 5.4). In this way, together

with the above analysis, it is suggested that the polypyrimidine tract is diffusely distributed in minimal introns from *Schistosoma mansoni*.

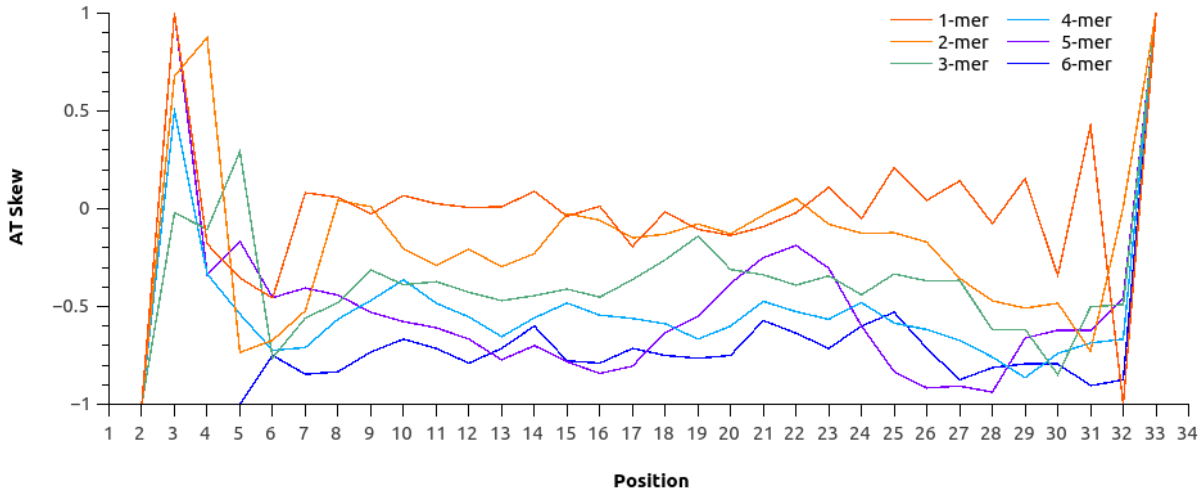


Figure 5.4 - AT skew of thymidines/adenosines tracts of different lengths. For each intronic position, the thymidine and adenosine were classified as being part of *k*-mer (1<=*k*<=6) of similar nucleotides. The AT skew of each k-mer tract was calculated based on the number of thymidine and adenosine by every position that belongs to the respective *k*-mer tract. Positive values indicate that more adenosine tracts than thymidine tracts were observed in the position. Negative values indicate the opposite. The analysis of AT skew for 34 bp introns (length of minimal intron peak, n=1,195) indicates that the presence of concatenation of the same type of nucleotide is higher for thymidines than for adenosines. The greater the *k*-mer analysed, the greater the tendency for thymidines.
Source: By the author.

5.3 Search for branch point consensus sequence of minimal introns

The splicing sites are highly conserved elements which mutations could lead to the reduction of the splicing efficiency or even its complete abolition. Branch point sequence, on the other hand, is not invariant. In fact, most metazoan introns lack perfect complementarity between the branch sequence and the U2 snRNA. Furthermore, the branch site utilization is affected by its surroundings. Deletions observed in polypyrimidine tract sequences, for example, may suppress lariat formation and thus splicing. [100] In this section, it will be discussed the different methods that we used for searching for the branch point consensus sequences of minimal introns from *Schistosoma mansoni*.

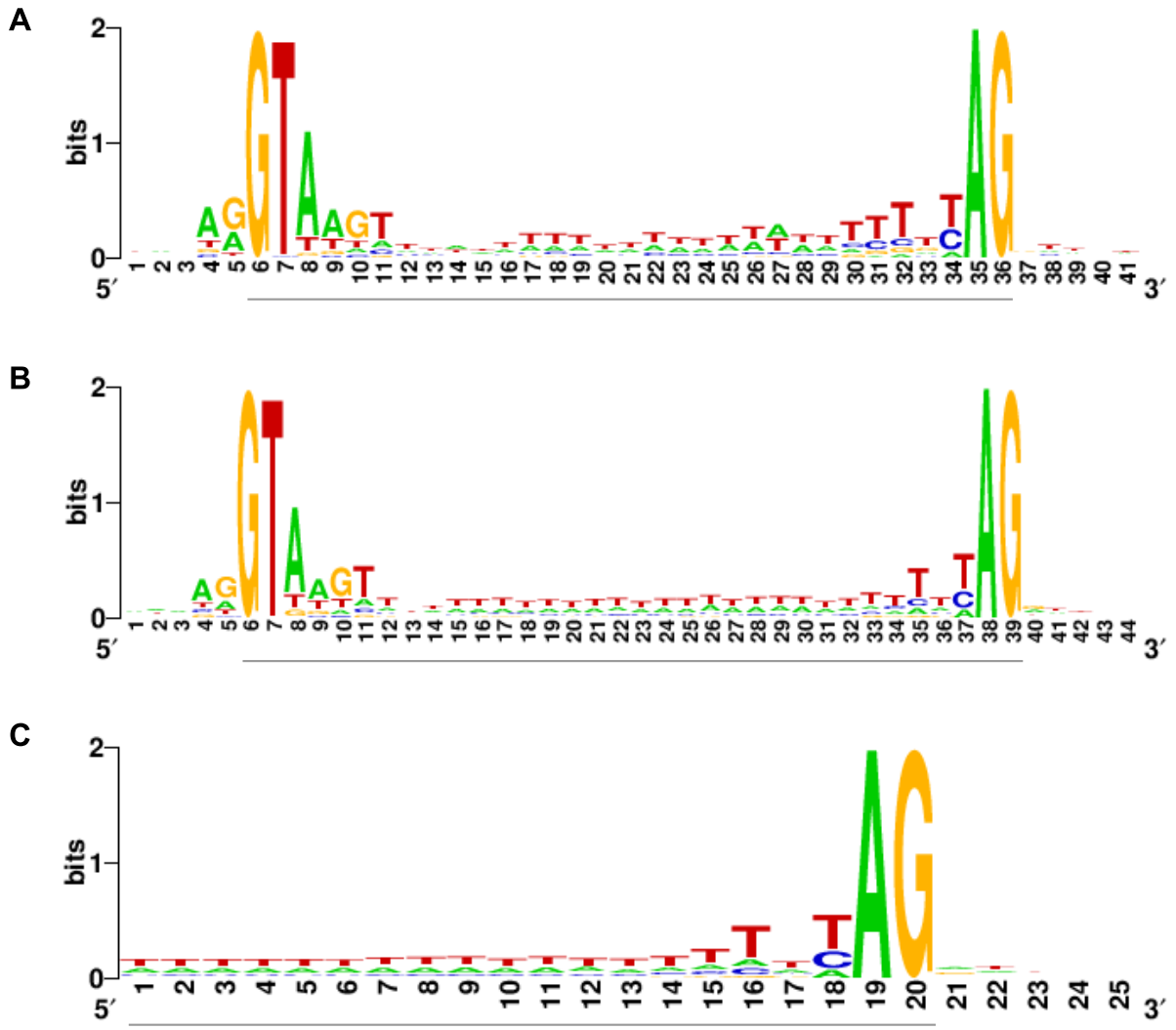


Figure 5.5 - Sequence conservation of nucleotides from minimal introns represented as a sequence logo. Height of nucleotide symbols indicates the relative frequency of the nucleotide in the respective position. Overall height of all symbols represents the sequence conservation at the position, measured in bits. Intronic positions are indicated at the bottom of the plots by a grey line. In (a) and (b), the sequence logo was created from the alignment of the entire sequences of 31 bp and 34 bp introns, respectively. A slight increase of A nucleotide at 8 nt upstream the 3' splice site was detected in (a). The same could not be observed in (b). In (c), the sequence logo was created with the alignment of the last 20 intronic positions from the entire population of minimal introns plus the first 5 nt of the respective downstream exon. No increase in A nucleotide was detected.

Source: By the author.

The branch sequence is characterized by the presence of an adenosine residue needed for the first transesterification reaction of the splicing process that results in the formation of the lariat intron structure. [30] Knowing the importance of the adenosine residue in the consensus sequence of the branch site, our first approach was to examine the nucleotide proportion and the information content along intronic positions. The branch point sequence is reported to show a flexible location within 19 nt to 37 nt

distance from the 3′ splice site in humans. [102] In yeast, the mean distance changes to 39 nt upstream the 3′ splice site. [103] As minimal introns of *Schistosoma mansoni* have a remarkable short length, we speculated that regulatory splicing elements would have a more restricted location and thus would be easily detected. In this way, this analysis was conducted using introns of 31 bp, the lower limit of minimal introns length, and introns of 34 bp, the length of the minimal intron peak. On 31 bp introns, a slight increase of adenosine residue at 8 nt upstream the 3′ splice site was detected (Figure 5.5a). However, this could not be replicated in the analysis of the 34 bp introns (Figure 5.5b). When applying the same methodology to the last 20 nt of all sizes of minimal introns plus 5 nt of the downstream exon, no preference of adenosine residue was detected (Figure 5.5c).

As a second approach for detecting the consensus sequences of the branch point, we decided to implement algorithms previously consolidated for DNA motif finding. W-AlignACE [104] and MEME [105] tools were used for motif detection of 31nt introns. Different search parameters such as motif wide and the number of motifs to find were used. However, no branch point consensus sequences could be detected. Usually, the determination of the branch point consensus sequence is achieved using both experimental and computational analysis. [106] In this way, based on the approach used for searching the branch point sequence, our results may not be conclusive to determine a specific position and sequence of the branch point of the minimal introns from *Schistosoma mansoni*.

5.4 Minimal introns have lower GC content than regular-sized introns

It was previously reported that shorter introns from the human genome have higher GC content than longer introns. Human introns between 50 bp to 86 bp show a median GC content of 63.53%. [80] Interestingly, based on our previous analysis, minimal introns from *Schistosoma mansoni* appear to be AT-rich. To verify if the same tendency observed in human introns could be applied to introns from *Schistosoma mansoni*, we calculated the GC% of minimal and regular-sized introns. Minimal introns

from *Schistosoma mansoni* have lower GC content than regular-sized introns (Figure 5.6). Interestingly, although results show a tendency for different GC content, minimal introns of both species show extremes in GC% when compared to other introns, which GC% is similar to the GC% of the rest of the genome.

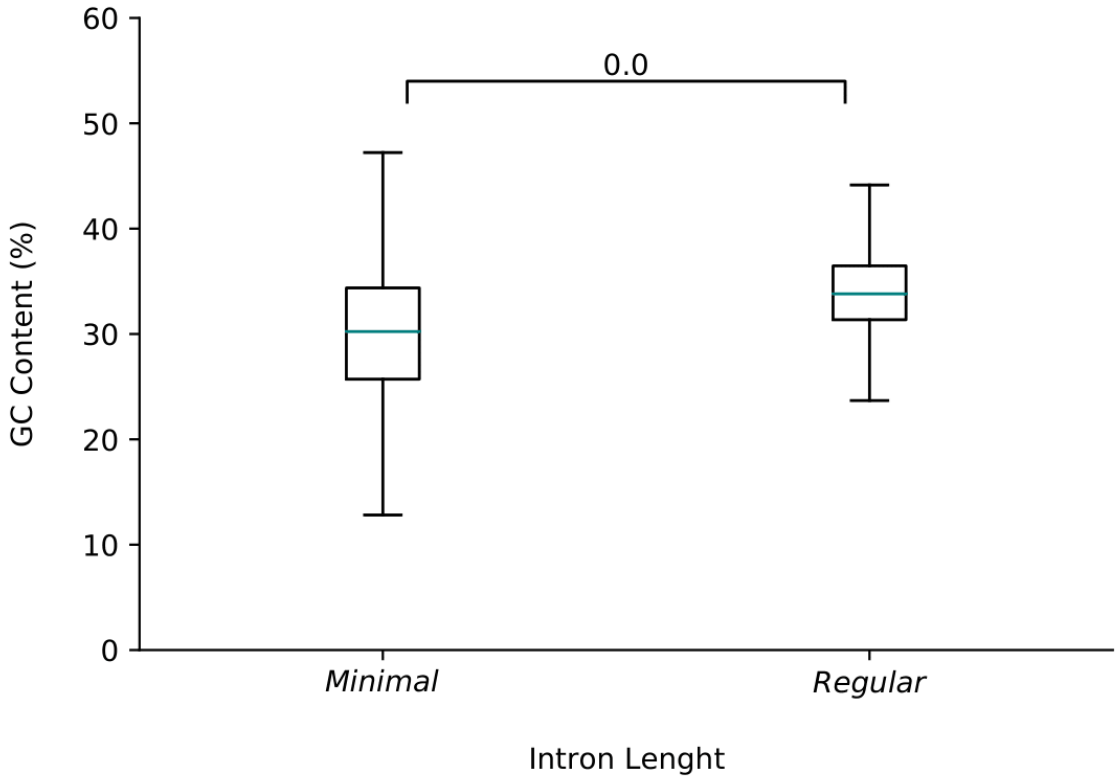


Figure 5.6 - GC content of minimal and regular-sized introns from *Schistosoma mansoni*. Minimal introns from *Schistosoma mansoni* have lower GC content than regular-sized introns. In the boxplot representation, the central line coloured in green represents the median, the first quartile (25th percentile) by the bottom line of the box, the third quartile (75th percentile) by the upper line of the box, and the maximum and minimum values by the lines at the end of the whiskers. P-value from the non-parametric Mann–Whitney U test is shown above the horizontal bracket.

Source: By the author.

5.5 Materials and methods

5.5.1 Nucleotide proportion along exon-intron boundaries

The frequency of each nucleotide was calculated for each position of the exon-intron boundaries of introns from the *Schistosoma mansoni* genome (version 7). Introns of 31 bp to 43 bp were classified as minimal (n=8,996) using our previous

definition based on minimal intron peak (see Chapter 4). Introns longer than 43 bp were classified as regular-sized introns (n=64,891). At the 5′ exon-intron boundary, 5 positions of the upstream exon and first 10 positions of the minimal intron were analysed. At the 3′ boundary, the last 10 intronic positions plus 5 positions of the downstream exon were used. The nucleotide proportion of minimal introns was calculated as an absolute value using all 8,996 introns. The nucleotide proportion of regular-sized introns, on the other hand, was calculated using randomly selected samples of 8,996 introns from the 64,891 population of regular-sized introns. The average nucleotide proportion of regular-sized introns and the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) was calculated from 10,000 simulations. For this and other tasks, Bash language (version 4.3.48) and Awk (version 4.1.3) text processing library were used and graphs were created using QtiPlot software (version 0.9.8.9), unless stated.

5.5.2 Information content along exon-intron boundaries

The information content (I) was calculated for each position (i) of the exon-intron boundaries found in the protein-coding genes from the *Schistosoma mansoni* genome (version 7). I can be defined as:

$$I_i = H_{before} - (H_{after} + e_n) ,$$

where H is the Shannon entropy [107] and e_n a correction factor of sample size (N):

$$H = -\sum_{k=1}^{n} p_k \, log_2 \, p_k$$

$$e_n = \frac{1}{ln2} \times \frac{n-1}{2N}$$

For DNA sequences, where only adenosine, guanosine, cytidine and thymidine nucleotides are possible, $n = 4$ and p_k is the probability/frequency to find the nucleotide k in the respective position. The log_2 found in the entropy formula is used

for obtaining values as bits as the unit of measurement. H_{before} is the entropy of the background probabilities of the nucleotide, and H_{after} is the entropy of the observed nucleotides probabilities. [108-109] For this analysis, it was considered that sequences have equal nucleotide background probability ($p_k = \frac{1}{4}$), therefore $H_{before} = 2\ bits$.

Introns of 31 bp to 43 bp were classified as minimal using our previous definition based on minimal intron peak (see Chapter 4). Introns longer than 43 bp were classified as regular-sized introns. At the 5′ exon-intron boundary, 5 positions of the upstream exon and first 10 positions of the minimal intron were analysed. At the 3′ boundary, the last 10 intronic positions plus 5 positions of the downstream exon were used. The information content of minimal introns was calculated as an absolute value using all 8,996 minimal introns. The information content of regular-sized introns, on the other hand, was calculated using randomly selected samples of 8,996 introns from the 64,891 population of regular-sized introns. The average information content of regular-sized introns and the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) was calculated from 10,000 simulations.

5.5.3 Thymidine proportion and information content within minimal introns

In this analysis, introns of 31 pb and 34 pb were analysed. The proportion of the thymidine nucleotide and the information content were calculated for each intronic position plus 5 positions of the adjacent exons as previously described. Furthermore, in order to avoid a possible nucleotide bias observed in the genome of the *Schistosoma mansoni*, a threshold curve of information content was calculated by randomly shuffling the nucleotides of each intronic sequence. The average information content of shuffled sequences and the confidence interval of 95% of confidence level (Z-score of 1.96 times the standard deviation) was calculated from 10,000 simulations.

5.5.4 Analysis of AT skew along the positions of minimal introns

In this analysis, introns of 34 pb (length of minimal intron peak, n=1,195) were analysed. Thymines and adenines were classified as being part of a subsequence of length *k* (1<=*k*<=6) of similar nucleotides. The AT skew of each *k*-mer tract was

calculated based on the number of thymine and adenine found at every intronic position that belongs to the respective *k*-mer tract. The AT skew of position i is defined as:

$$ATskew_i = \frac{c_A - c_T}{c_A + c_T}$$

where c_A and c_T represent the number of adenines and thymines found at the position. [110] Positive values indicate that more adenine tracts than thymine tracts were observed in the position i. Negative values indicate the opposite.

5.5.5 Sequence logo of minimal introns

Entire sequences of minimal introns of 31 bp (the lower limit of minimal introns length) and 34 bp (the length of the minimal intron peak) plus 5 positions of the adjacent exons, and sequences of last 20 positions plus 5 positions of the downstream exon from all minimal introns were analysed using sequences logos. The sequence alignment, nucleotide proportion, information content and graph creation of each analysis were all computed by the WebLogo tool. [111] In the graphs, the height of nucleotide symbols indicates the relative frequency of the nucleotide in the respective position. Overall height of all symbols represents the sequence conservation at the position, measured in bits. [111]

5.5.6 Searching for branch point consensus sequence with motif finding algorithms

In order to search for the branch point consensus sequences of minimal introns from *Schistosoma mansoni*, W-AlignACE [104] and MEME [105] tools were used for motif detection of 31 bp introns. Both algorithms are based on probabilistic methods. However, W-AlignACE uses a stochastic approach based on Gibbs' sampling, whereas MEME has a deterministic approach based on the expectation-maximization concept. [112] For analysis using the W-AlignACE tool, the number of columns to align was set as 10, the number of sites to expect was 10 and the fractional background GC content was set as 0.38. Three analyses were conducted using the MEME tool. In the

first attempt, we set as 3 the number of motifs to find, with a minimum width of 2 nt and maximum of 10 nt. In the second attempt, we set as 4 the number of motifs to find, with a width of 7 nt. In the third attempt, we set as 5 the number of motifs to find, with a minimum width of 4 nt and maximum of 15 nt. In the first and second attempts, motifs were searched in the given strand only. For all attempts, the other parameters were left in the default mode: classic mode of motif discovery and distribution sites of any number of repetitions.

5.5.7 GC content of introns from *Schistosoma mansoni*

The GC content of minimal and regular sized introns of *Schistosoma mansoni* was computed using Bash language (version 4.3.48) and Awk (version 4.1.3) text processing library. Introns of 31 bp to 43 bp were classified as minimal (n=8,996) using our previous definition based on minimal intron peak (see Chapter 4). Introns longer than 43 bp were classified as regular-sized introns (n=64,891). Boxplot graph and Mann-Whitney U statistical test were done in Python programming language (version 3.7.3) with support of Pandas (version 0.23.4), Matplotlib (version 2.2.3), Seaborn (version 0.9.0) and Scipy (version 1.3.1) libraries.

Chapter 6

INTRON RETENTION OF MINIMAL INTRONS FROM *SCHISTOSOMA MANSONI*

Gene expression in eukaryotes is known for having multiple integrated and highly regulated steps. A series of mRNA processing is essential before proper translation. As a consequence of this complex pathway, eukaryotes are able to achieve great levels of genome plasticity. As a counterpart, the existence of many steps in transcription processing increases the probability of errors. To ensure the viability of such a system, eukaryotes have also developed many surveillance mechanisms in the cell. The NMD pathway is one of those mechanisms that have been most studied and appears across the eukaryotic lineage .[113]

The NMD is responsible for a rapid decay of erroneous transcripts that harbours a PTC on its sequence. The introduction of a PTC may be a result of splicing mistakes such as intron retention which translation could result in the expression of aberrant proteins. But how does the cell discriminate a PTC from a proper stop codon? The recognition of a PTC is dependent on the translation step, where the crosstalk of the ribosome with downstream *cis*-acting elements of the mRNA leads to the recruitment of *trans*-acting NMD factors and thus mRNA degradation. These *cis*-acting elements, however, varies between different species. In mammals, for example, the recruitment of NMD factors is linked to the crosstalk between the ribosome and a splicing-generated EJC. [114] In contrast, many other analysed species have shown that NMD activation is independent of EJC, including *Saccharomyces cerevisiae*, [115] *Drosophila melanogaster*, [116] *Tetrahymena thermophila* [117] and *Caenorhabditis elegans*. [118]

Interestingly, it is well known that splicing processes may occur either via intron or exon definition. [28] Long introns are commonly removed via exon definition. In this

case, an error in the splicing process leads to exon skipping. On the other hand, splicing of short size introns is accomplished via intron definition, for which errors lead to intron retention. [66-67] Due to the known short lengths of minimal introns, it is suggested that this type of introns is governed by the intron definition fashion of splicing. In this chapter, we investigate if retention of minimal introns from *Schistosoma mansoni* is under selection for the introduction of PTC in the transcripts that may reduce its deleterious effect. Here, we show that the preference for location and composition of PTC in minimal introns shows evidence of selection processes. Furthermore, we show that the percentage of PTC-containing introns is greater in minimal introns whose size is multiple of 3, in which retention does not shift the reading frame of translation.

6.1 Splice sites contribute to PTC formation

In this and the following analysis, the rationale for studying intron retention includes only minimal introns located in the coding region of the mRNA and consider only one intron retention event at a time. TAA, TGA and TAG nonsense codons are considered as PTCs. Introns in which PTC includes at least one intronic nucleotide are said to be PTC-containing introns. PTC location inside introns is measured by numbers of intronic codons. The introduction of a PTC takes into account the reading frame in case of retention. Is classified as phase 0, introns whose first codon starts in the first intronic position. As phase 1, if the second codon is started at the second intronic position. As phase 2, if the second codon starts at the third intronic position. In this way, the first codon of introns at phase 1 and 2 is partially composed by exonic nucleotides. Depending on the size and phase of the intron, the last codon may also include nucleotides from the downstream exon.

In a previous study, it was demonstrated that PTC is selected to occur earlier than expected by chance in introns of seven model organisms. Authors claim that PTC location in the intron is due to selection for reducing waste and increasing the efficiency of degradation of aberrant transcripts. [119] Our investigation regarding the PTC introduction by retention of minimal introns of *Schistosoma mansoni* shows that PTCs

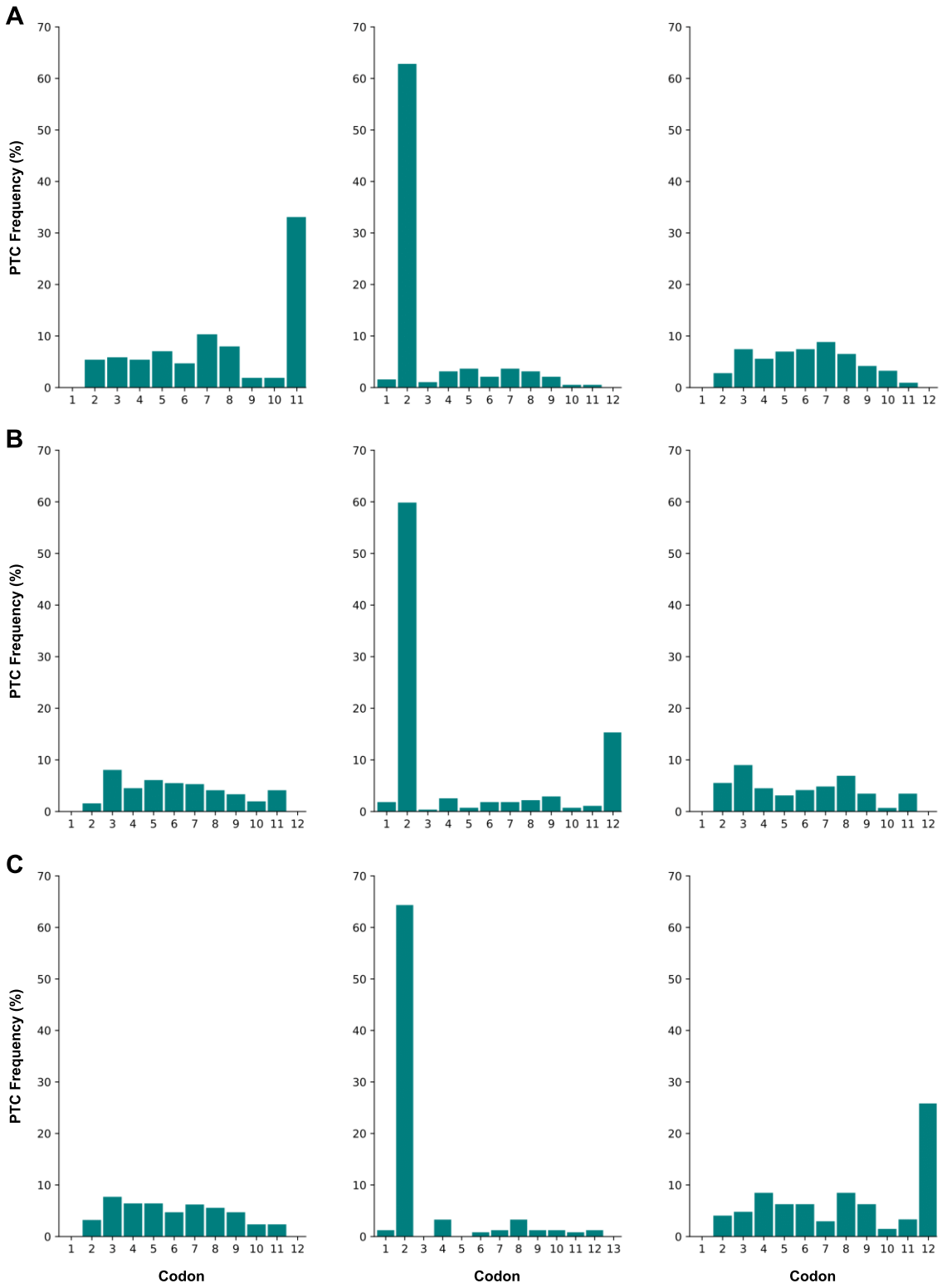


Figure 6.1 - Location of PTC along intronic codons. The percentage of each PTC position was computed for every size of minimal intron and phase of the reading frame taking into account introns with and without nonsense codons. Only graphs of introns with 33 bp (a), 34 bp (b) and 35 bp (c) are shown as a representation to the results seen in introns which size is multiple of 3 (3n), multiple of 3 plus 1 (3n+1) or multiple of 3 plus 2 (3n+2), respectively. Introns whose retention occurs at phase 0 are shown in the left, phase 1 in the middle and phase 2 in the right of the figure. PTC are preferentially located in the second and last codons of the intron. These preferences are dependent on introns phase if at second codon or dependent on both intron size and phase if at last codon.

Source: By the author.

are more prevalent at the second and last codons of the intron (Figure 6.1). This prevalence of the PTC in the second codon, however, only occurs at introns at phase 1. Minimal introns with PTC at last codon position are composed exclusively by introns on phase 1, 2 and 0 if introns size is multiple of 3 (3n introns), multiple of 3 plus 1 (3n+1) or multiple of 3 plus 2 (3n+2), respectively (Table 6.1).

Table 6.1 - Percentage of phase of reading frame in case of retention of different size groups of minimal introns and PTC location.

Codon of PTC	Intron Size Group	Phase 0 (%)	Phase 1 (%)	Phase 2 (%)
2nd	3n	12.12	80.56	7.32
2nd	3n+1	6.43	85.48	8.09
2nd	3n+2	7.64	85.59	6.77
Last	3n	100	0	0
Last	3n+1	0	100	0
Last	3n+2	0	0	100

Source: By the author.

If analysing the composition of the PTC at second and last codons taking into account the described dependence on introns size and phase, it is shown that the splice sites are important elements for its formation. The PTCs at second codon are mainly composed by the 2nd-4th nucleotides of the intron as minimal introns with PTC at second codon position are enriched in phase 1. In other words, it is partially composed by the 5′ splice site GT (Figure 6.2a). The PTC at last codon includes the AG of the 3′ splice site in all cases as there is 100% of phase 0, 1 and 2 of minimal introns of size 3n, 3n+1 and 3n+2, respectively. Thus, the only possible nonsense codon would be the TAG (Figure 6.2b). The phases in which the splice sites nucleotides are part of the PTC are the exactly enriched phases of the PTC-containing introns at second and last codon positions.

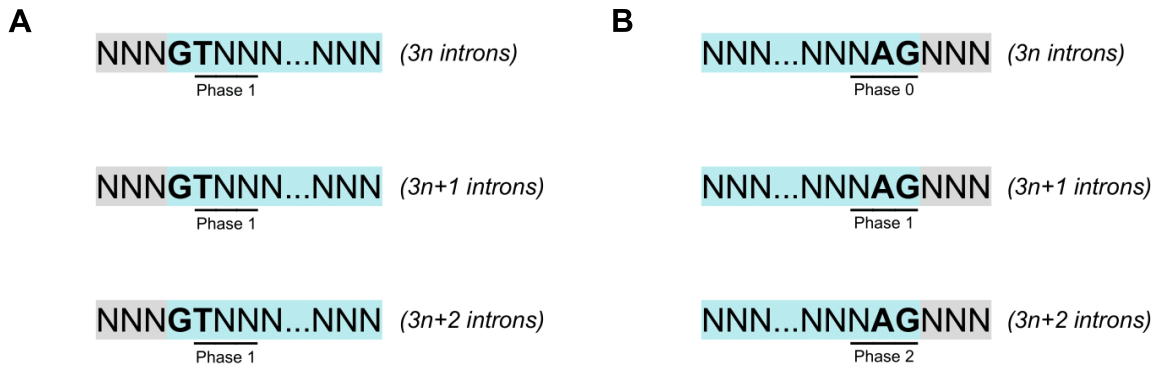


Figure 6.2 - Schematic representation of the contribution of splice sites to the formation of PTC in case of retention of minimal introns. The 5′ splice site may contribute with the T nucleotide to the formation of the nonsense codons TAA, TGA and TAG located at the second intronic codon. For all sizes of minimal introns, this contribution may occur if the intron is at phase 1 (a). The 3′ splice site could only contribute to the formation of the TAG nonsense codon located at the last codon of the intron. The occurrence of such a contribution is dependent on the size and phase of the intron (b). The phases in which the splice sites nucleotides are part of the PTC are the exactly enriched phases of the PTC-containing introns at second and last codon positions.
Source: By the author.

Furthermore, analysis of sequence logo of minimal introns with PTC at second codon position shows an increase of A and G nucleotides at the 8th and 9th intronic nucleotide positions (Figure 6.3a). Those nucleotides, together with contribution of the T nucleotide of the 5′ splice site, may compose nonsense codons. On the other hand, sequence logo of minimal introns in which PTC is at last codon position shows an increase of T nucleotide at those 8th and 9th intronic positions (Figure 6.3b). The T nucleotide on those positions disrupts the possibility of PTC creation using the T nucleotide of the 5′ splice site. The different selection for nucleotides also varies for the antepenult intronic nucleotide of minimal introns harbouring the PTC at second or last codons positions. The proportion of the T nucleotide at the antepenult intronic position is absolute when the PTC is at the last codon position as the only possible nonsense codon is TAG due to the AG contribution of the 3′ splice site. The proportion of T nucleotides, however, is not such prominent in minimal introns with PTC at second codon position. Together, these results show the contribution of the 5′ and also 3′ splice sites to the introduction of PTC in case of retention of minimal introns.

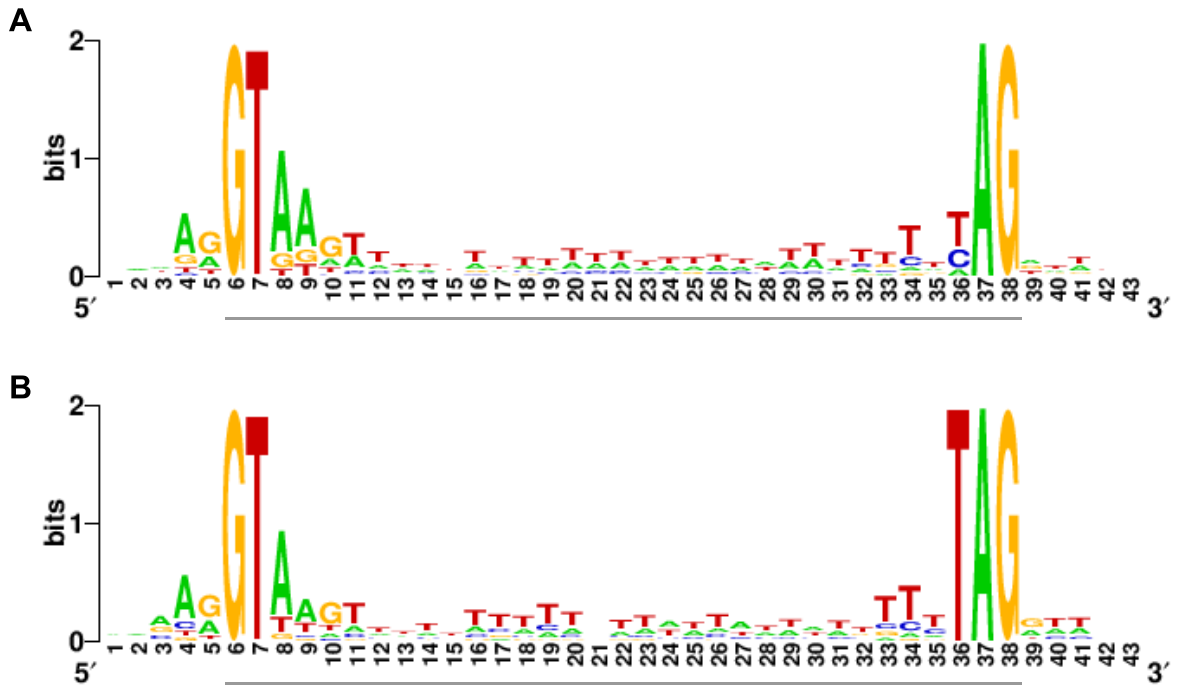


Figure 6.3 - Sequence logo of minimal introns with PTC. Sequence logos were created for every size of minimal introns plus 5nt of the adjacent exons. Only logos from introns with 33 bp are shown. Intronic positions are indicated at the bottom of the plots by a grey line. Height of nucleotide symbols indicates the relative frequency of the nucleotide in the respective position. Overall height of all symbols represents the sequence conservation at the position, measured in bits. Sequence logos were created from the alignment of the sequences of 33 bp introns which PTC is located at the second (a) and last (b) intronic codon positions. In total, 149 introns were aligned in (a) and 141 in (b). The PTCs at second codon are mainly composed by the 2nd-4th nucleotides of the intron, it is partially composed by the 5′ splice site GT (a). The PTC at last codon includes the AG of the 3′ splice site in all cases and thus the only possible nonsense codon would be the TAG (b). The selection for nucleotides in the 3rd, 4th and antepenult intronic positions are subject to the position of the PTC.

Source: By the author.

6.2 Symmetric minimal introns are selected to harbour PTC

Introduction of PTC by intron retention could be caused by the presence of a nonsense codon in the retained intron or it can be a product of a shift of the reading frame from the exons downstream of a non-symmetric retained intron. Symmetric introns (3n introns), however, do not shift the reading frame in case of retention as codons are defined as nucleotide triplets. In this way, our next question is: does the selection for intronic sequences that results in nonsense codons is stronger in symmetric introns? If positive, we should see an increase in the percentage of introns

in which PTC is found inside intronic sequences within 3n introns, as a PTC could not be created downstream the intron sequence. Interestingly, when computing the percentage of PTC-containing introns of all sizes of minimal introns, 3n introns (33 bp, 36 bp, 39 bp and 42bp) show a higher percentage than other size of introns (Figure 6.4). Curiously, 3n+2 minimal introns (32 bp, 35 bp, 38 bp and 41 bp) display an intermediate frequency of PTC-containing introns when compared to 3n and 3n+1 introns. Furthermore, the overall percentage of PTC-containing introns increase if longer the minimal intron for 3n+1 and 3n+2 introns only, suggesting a different dynamic to 3n introns.

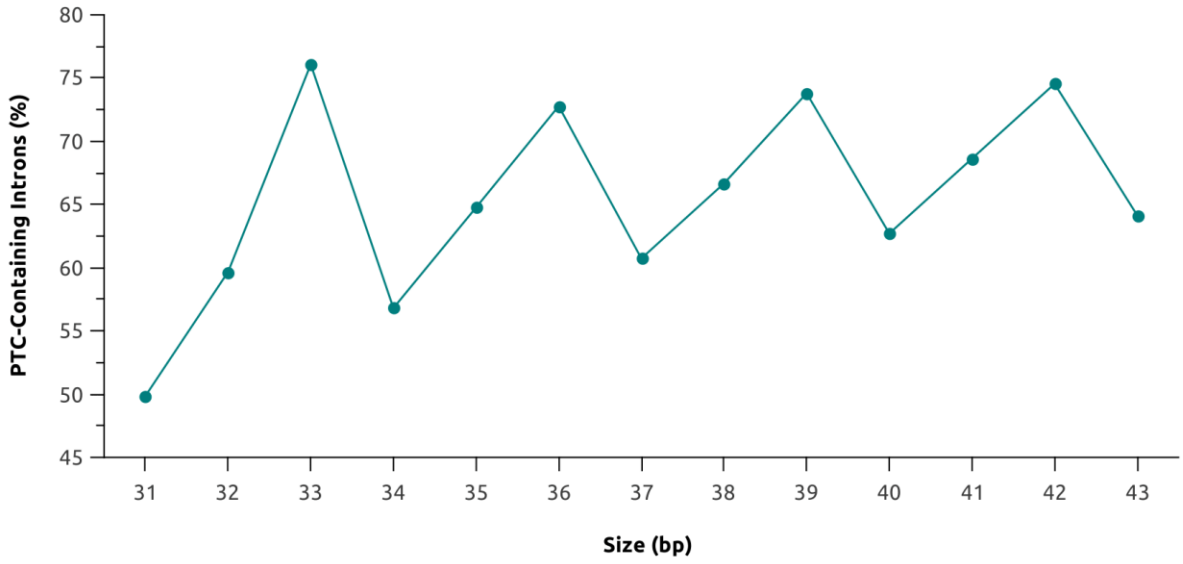


Figure 6.4 - Percentage of PTC-containing introns per size of minimal introns. From the total of introns found in each size, the percentages of introns with PTC in case of intron retention are shown. Introns whose size is multiple of 3 (3n) shows a higher percentage than other size introns. Introns whose size is multiple of 3 plus 2 (3n+2) have intermediate values of percentage.
Source: By the author.

In order to further investigate why 3n minimal introns have a higher level of PTC-containing introns and why 3n+2 minimal introns display intermediate values, we also analysed the percentage of the phase of reading frame in case of intron retention. Almost half of minimal introns are classified as phase 0 (phase 0 = 47.88%, phase 1 = 23.14%, phase 2 = 28.97%). Together with our previous observation showing that the contribution of splice sites on PTC formation is dependent on introns size and phase

of retention, the variation of percentage of PTC-containing introns observed in different sizes of minimal introns could be explained. We propose that 3n introns show a higher proportion of PTC because it is the only type of intron that allows the formation of a PTC at the last codon position in phase 0 (Figure 6.2). Introns retained at phase 1 of the reading frame could introduce a PTC composed of the 5′ splice site nucleotide regardless of the size of the retained intron. At phase 2, however, only 3n+2 introns have a PTC with the contribution of the 3′ splice site. In this way, the intermediate frequencies of PTC-containing introns by 3n+2 introns may be explained as both phase 1 and 2 favours the contribution of the splice sites in PTC formation. On the other hand, 3n+1 introns just show PTC composed by 5′ and 3′ splice sites if at phase 1, the same phase that all sizes of introns have a PTC formation facilitated by the use of the 5′ splice site.

6.3 PTC-containing introns are randomly distributed in the gene

In Chapter 4 “Distribution of minimal introns in the genome of *Schistosoma mansoni*”, we have discussed that minimal introns of *Schistosoma mansoni* are predominant towards the 5′ end of the gene. In this new analysis, we investigated if there is also a predominance in the distribution of minimal introns along the gene in terms of its size and percentage of PTC-containing introns. To overcome the bias of minimal introns to accumulate in the 5′ end of the gene, the results are present as a fraction of the analysed parameter concerning the total number of minimal introns found in the respective gene position. Positions in the gene are counted in relation to the 5′ end. Genes were divided in relation to the number of harbouring introns, so genes with only one intron do not enrich our result towards the first intronic position.

Firstly, the distribution of the size of minimal introns in the gene was evaluated. The fractions of 3n, 3n+1 and 3n+2 introns were analysed. All minimal introns were considered, including introns with and without PTC in case of intron retention. As shown in Figure 6.5, no tendency towards any gene position was observed. Later, we applied the same methodology to obtain the fraction of the minimal introns with PTC in

relation to all minimal introns. Once more, no preference for gene position was observed (Figure 6.6).

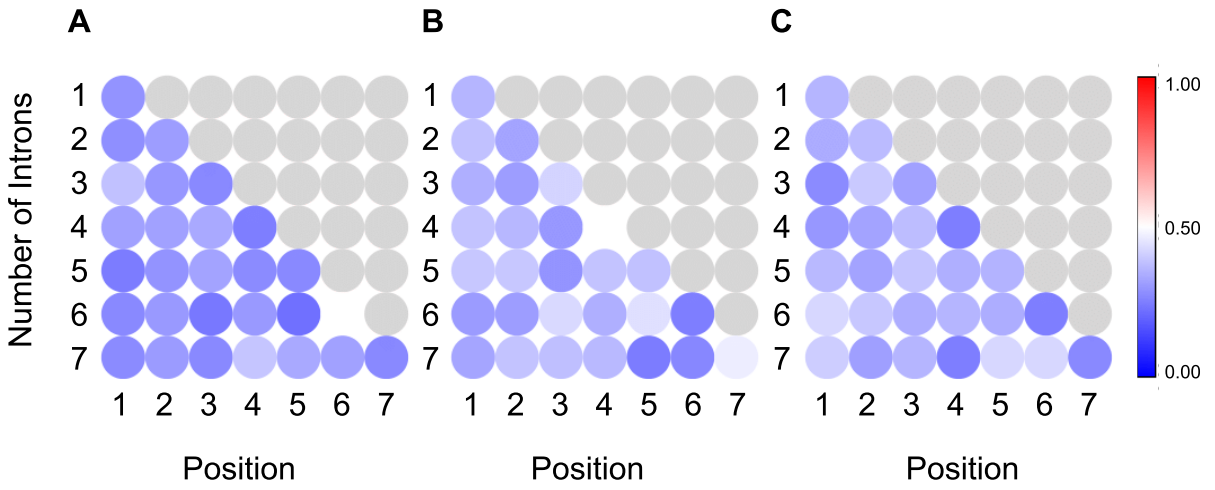


Figure 6.5 - Distribution of size of minimal introns along the gene. Genes were divided in relation to the number of harbouring introns in the coding region. Positions in the gene are counted in relation to the 5′ end. Distribution of introns whose size is multiple of 3 (3n introns), multiple of 3 plus 1 (3n+1) or multiple of 3 plus 2 (3n+2) is shown in (a), (b) and (c), respectively. The fraction of the intron group in relation to the total number of minimal introns found in the respective gene position is shown by a coloured scale. Columns represent intron positions along the gene and rows represent the number of introns in the coding region of the analysed genes.

Source: By the author.

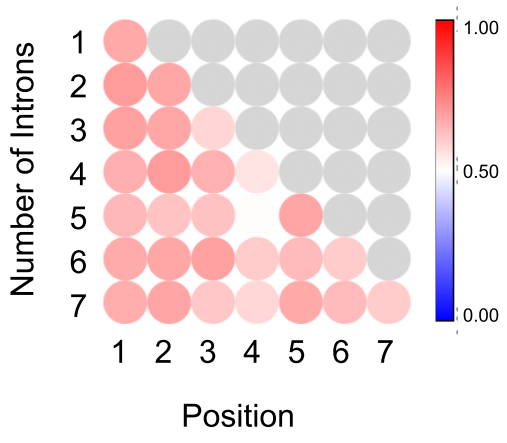


Figure 6.6 - Distribution of minimal introns with PTC along the gene. Genes were divided in relation to the number of harbouring introns in the CDS. Positions in the gene are counted in relation to the 5′ end. The fraction of the minimal introns with PTC in relation to the total number of minimal introns found in the respective gene position is shown by a coloured scale. Columns represent intron positions along the gene and rows represent the number of introns in the coding region of the analysed genes.

Source: By the author.

6.4 Materials and methods

6.4.1 Intron coordinates in case of retention

Original coordinates of minimal introns (introns between 31 to 43 bp) were obtained according to the methodology described in Chapter 4 “Distribution of minimal introns in the genome of *Schistosoma mansoni*”. Knowing the extension of the UTR described in the genome annotation file, minimal introns were divided into those located at UTR and coding regions. For this and the following analysis, only minimal introns at coding regions were considered for the study. The phase of minimal introns in case of retention was determined based on the phase of the immediately downstream CDS described in the genome annotation file. The same phase attributed to the downstream CDS was applied to the respective minimal intron. Sequences of introns were later divided into codons (a subsequence of triplet nucleotides). If the reading frame of the retained intron is at phase 0, the first intronic codon comprises the first three intronic nucleotides. Otherwise, if at phase 1, the first codon includes the last two nucleotides of the upstream exon plus the first intronic nucleotide. And if at phase 2, the first codon consists of the last exonic nucleotide plus the first two intronic nucleotides. The composition of the last intronic codon depends on the size of the minimal intron and the phase of the reading frame in case of retention. Last codons may include the presence of nucleotides of the downstream exon. In this way, coordinates of introns were calculated in the case of independent retention. In other words, if only one intron is retained in the transcript at a time. For this task, Bash programming language (version 4.3.48) and Awk text processing library (version 4.1.3) were used.

6.4.2 Location of PTC in the intron

As previously described, minimal introns were separated in terms of size and the phase of the reading frame in case of retention. Screening all the codons of the retained introns, the position of the first nonsense codon defined as TAA, TGA or TAG were computed. This codon is named PTC. The percentage of PTC of each group of intron size (31 bp to 43 bp introns) and phase (phases 0, 1 or 2) was calculated for

every codon position. In this way, if there are 100 introns of 31 bp at phase 1 (includes intron with and without PTC), and 30 of those introns show the PTC at the last codon position, there will be 30% of PTC at the last codon position. Be aware that the number of codons depends not only in the size of the intron but also on its phase. In this example, the number of codons is 11. The equations showing how the number of codons is calculated are shown in Table 6.2. For this task, Bash programming language (version 4.3.48) and Awk text processing library (version 4.1.3) were used. Plots were created using Python programming language (version 3.7.3) with support of Matplotlib (version 2.2.3) library.

Table 6.2 - The number of codons is dependent on the intron size and phase of its reading frame.

Intron Size (s)	Phase of Reading Frame	Number of Codons
3n	0	$s\ /\ 3$
3n	1	$(s+3)\ /\ 3$
3n	2	$(s+3)\ /\ 3$
3n+1	0	$(s+2)\ /\ 3$
3n+1	1	$(s+2)\ /\ 3$
3n+1	2	$(s+2)\ /\ 3$
3n+2	0	$(s+1)\ /\ 3$
3n+2	1	$(s+4)\ /\ 3$
3n+2	2	$(s+1)\ /\ 3$

Source: By the author.

6.4.3 Percentage of the phase of reading frame in case of intron retention

After determining the phase of the reading frame of introns in case of retention explained in section 6.4.1, “Intron coordinates in case of retention", the percentage of each phase for many retention situations were analysed. Firstly, the percentage of each phase was calculated in relation to all minimal introns, from 31 bp to 43 bp, with or without nonsense codons. Then, minimal introns which PTC is located at the second codon position were grouped according to their length as: 3n (multiple of 3), 3n+1, and

3n+2. Percentage of each phase was calculated for each group. In this way, if there are 100 minimal introns classified as 3n and with the PTC at the second codon position and 12 of these introns is at phase 0, 80 at phase 1 and 8 at phase 2, then the percentage of each phase would be 12%, 80% and 8% respectively. This same calculation was also done to all groups of minimal intron sizes with a PTC located at the last codon position. This analysis was conducted using Bash programming language (version 4.3.48) and Awk text processing library (version 4.1.3).

6.4.4 Sequence logo of minimal introns with PTC at second or last codons

Minimal introns with PTC located at the second or last codons were divided according to its intron length, from 31 bp to 43 bp. For each group, the sequence of the entire intron plus five nucleotides of the adjacent exons were extracted. A sequence logo was done taking as input a fasta file of all sequences for each group. This analysis was conducted using Bash programming language (version 4.3.48) and Awk text processing library (version 4.1.3). Sequence logos were created using the WebLogo tool. [111]

6.4.5 Percentage of PTC-containing introns for each size of minimal introns

The percentage of introns with a PTC in case of retention was computed for each size of minimal introns, from 31 bp to 43 bp. This task was conducted using Bash programming language (version 4.3.48) and Awk text processing library (version 4.1.3). Graphs were created using QtiPlot software (version 0.9.8.9).

6.4.6 Distribution of minimal introns along the gene

Genes were grouped according to the number of coding region introns. For each group of genes, the number of minimal introns was computed for each intronic position. In the first analysis, the number of minimal introns of length 3n, 3n+1 and 3n+2 was also obtained for each intronic position. The percentage of each group of intron length was calculated using the previous number of minimal introns in the same analysed location. On this way, for genes with three harbouring introns, if the first position has 200 minimal introns, the second has 100 minimal introns and the third has 50 minimal

introns, of this 100 minimal introns are classified as 3n in the first position, 40 at the second position and 10 at the third position, then the percentage of 3n minimal introns will be 50%, 40%, and 20% for the first, second and third intronic position respectively. In the second analysis, the number of minimal introns containing PTC was computed for each intronic position. The percentage of each group of intron length was calculated using the previous number of minimal introns in the same analysed location. For both analyses, genes with more than 7 introns per gene were not considered as the number of minimal introns in the last intronic position is too low. From the 483 genes with 8 coding region introns, only 8 introns were classified as minimal in the last intronic position of the gene. These analyses were conducted using Python programming language (version 3.7.3) with the support of Pandas (version 0.23.4) library for data manipulation. Graphs were created using Morpheus software for matrix visualization.
120

Chapter 7

CONCLUSION

The existence of minimal introns in many eukaryotic species and the selective pressure for maintenance of their optimal size have made scientists speculate about their functions and the unique features shared between genes that harbour them. In fact, minimal intron-containing genes of human and mouse genomes have been shown to be enriched with house-keeping functions, to be localized in the vicinity of open chromatin and to be more efficiently exported from the nucleus to the cytoplasm. [82] Together, these observations corroborate to the elaboration of the “Routing hypothesis”, where minimal intron-containing genes would have a particular chromosome location and recognition that facilitates their transcript exportation to the cytoplasm. [77,82] Interestingly, our analysis of minimal introns from *Schistosoma mansoni* have shown that those introns are enriched in some chromosomes. This result is consistent with what was previously reported in the human genome. [82] More analysis should be done in order to understand the meaning and consequences of this observed tendency.

Although eukaryotes share the existence of minimal introns in their genomes, each species has a particular minimal intron size. It was previously suggested that the modal size value of the peak would be a consequence of the differences in the splicing machinery observed in each species. [77] Our study with *Schistosoma mansoni* has shown that minimal introns are concentrated around the size of 34 bp, a relatively low number when compared to what is commonly observed in other eukaryotic species. The analysis of the distribution of minimal introns lengths has shown an intriguing sharp decline in the frequency of introns shorter than the modal length. We suggest that this asymmetry in the minimal intron peak may reflect a strong negative selection for too short size introns that could be linked to the physical constraints imposed by the

splicing machinery. If this is true, the physical constraints of the splicing machinery between species would vary greatly and further studies are warranted to understand what are the proteins determining this size.

Despite the fact that minimal introns display a size near the minimum possible for the occurrence of splicing, their splicing recognition signals do not appear to be much different to that from regular-sized introns. However, the information content of intron-exon boundaries was higher for regular-sized introns at some points. Interestingly, this observation was also reported in introns from *Caenorhabditis elegans*, where splice sites of introns longer than 75 bp have greater information content than shorter introns. [121] We speculate that this higher content results from the fact that regular-sized introns would have a higher probability to create spurious weaker recognition sites and therefore the true sites should be stronger to allow distinction from those weak spurious sites. In minimal introns, on the other hand, the chances of false positives marks are extremely reduced.

Curiously, we verified that the canonical splice sites GT and AG at the extremities of minimal introns from *Schistosoma mansoni* may have an important role in facilitating PTC introduction in case of intron retention. We showed that the majority of PTCs in minimal introns includes nucleotides from those splice sites. The fraction of PTC introduction is even higher in symmetric minimal introns, where retention does not shift the reading frame of translation, thus suggesting a biological role. Furthermore, symmetric minimal introns do not show an increase in the fraction of introns with PTC as the intron gets longer as is observed in 3n+1 and 3n+2 minimal introns and as would be expected in a stochastic phenomenon. We speculate that the intron retention in a fraction of genes containing those symmetric introns could have the function of producing actual protein isoforms with a few additional amino acids residues and thus they would be selected for not having nonsense codons inside their sequence. That same role cannot be fulfilled by non-symmetrical introns since their retention would introduce frameshifts.

In addition, the PTC-containing introns in *Schistosoma mansoni* are equally distributed in the genes, with no preference for concentration in the 5′ or 3′ regions. This contrast to the results of studies with mammals genomes, where PTCs were more

scarce near the 3′ end of the gene since PTCs located 50 nt or less from to the last EJC does not activate the NMD pathway as the ribosome is able to displace all EJCs present in the transcript. [122] As minimal introns from *Schistosoma mansoni* are shorter than 50 bp, the presence of PTC in any intronic codon could not activate NMD if they are found at the last position in the gene. We suggest that maybe the crosstalk between the ribosome and *cis*-acting elements that leads to NMD activation in *Schistosoma mansoni* may be independent of the EJC as observed in *Saccharomyces cerevisiae*, [115] *Drosophila melanogaster*, [116] *Tetrahymena thermophila* [117] and *Caenorhabditis elegans.* [118]

Part III

MINIMAL INTRONS OF VERTEBRATES

The Vertebrata is a phylum of the triploblastic bilaterian animals, classified as deuterostomes in which the blastopore gives rise to the anus. This phylum is part of the Chordata clade, with the Urochordata as its sister group that together comprises the Olfactores characterized by similarities in pharyngeal re-modification that lead to the formation of new structures. [123] The Vertebrata includes approximately 58,000 living species that comprises seven clades, which divergence has been related to change in the atmospheric oxygen and pressures affecting prey in the latest Precambrian and in the Palaeozoic eras respectively (~600-360 Ma). [124] Along the unique features shared by vertebrates, we can cite the neural crest, an endoskeleton, an adaptive immune system, and the genome constitution. [123] The uniqueness of the vertebrate genome is evident from analysis of genome-wide functional diversification of genes across genomes of metazoans. [125] It is argued that the vertebrate genome has experienced two rounds of genome-wide duplication (known as 2RGD hypothesis) that leads to increased genome complexity in those animals. [126-127] This event is expected to have occurred in the lineage leading to vertebrates and could be related to the observed increased morphological complexity under developmental control of this phylum. [123] Interestingly, the increase in complexity of vertebrates has also been correlated to expansion in intron length during its evolution, especially when it comes to mammals. These longer introns are reported to influence the splicing in many ways resulting in a more intricated transcriptome. [128]

As vertebrates have experienced an increase in intron lengths, some have argued that introns that need to remain short have been selected from those that do not. [34] In this section, we have chosen the vertebrates as a system for studying the evolution of minimal introns and understanding its persistence in those genomes that have been subjected to a process of intron expansion. In the first chapter, we begin by analysing minimal introns across the clades of the Deuterostomia group in order to have a broad look into the minimal introns of vertebrates in relation to their closest relatives. Later, we related how changes during vertebrate evolution have influenced

minimal introns, especially the homeothermy acquisition. Then, taking the human genome as a model of our study, a more systematic and detailed view relating minimal introns to intrinsic cellular functions is discussed. In the last chapter, “Conclusion”, we stated our main new findings of vertebrate minimal introns and how our results may impact the understanding of the evolution of the vertebrates genomes and morphologies.

Chapter 8

THE EVOLUTION OF INTRON SIZE IN DEUTEROSTOMES

Minimal introns are referred to by a sharp peak in the intron size distribution plots that represent a frequent accumulation of introns near the minimum size. They can be detected throughout the eukaryotic evolution, from unicellular to multicellular species. Protozoa, fungi, nematodes, arthropods, plants and many vertebrates' representatives have shown the persistence of such minimal introns across and within lineages. [34,77] However, many more species should have their genomes analysed for a better perception of the evolution of the intron size distribution. In this way, except the Vertebrata phylum, the deuterostome animals are still poorly understood. From the literature review, no data of minimal introns from Echinodermata, Hemichordata, Cephalochordata and Urochordata phyla of the Deuterostomia group was found. In this chapter, we aim to further collaborate with the study of the intron size evolution by investigating the introns from the deuterostomes as we believe that this knowledge might represent a good source for better understanding the architecture of modern genomes.

8.1 Many deuterostome groups do not show the presence of minimal introns

After studying minimal introns from the Platyhelminth *Schistosoma mansoni*, we decided to investigate how intron size is distributed throughout the Animalia kingdom, especially in the Deuterostomia clade, which has been poorly explored. The deuterostome animals differ from other bilaterian species by many developmental features, such as the fate of the blastopore from where its name is derived. Representatives from each deuterostome phyla, Echinodermata, Hemichordata,

Cephalochordata, Urochordata and Vertebrata, were selected. The chosen species and their phylogenic relationship are represented in Figure 8.1.

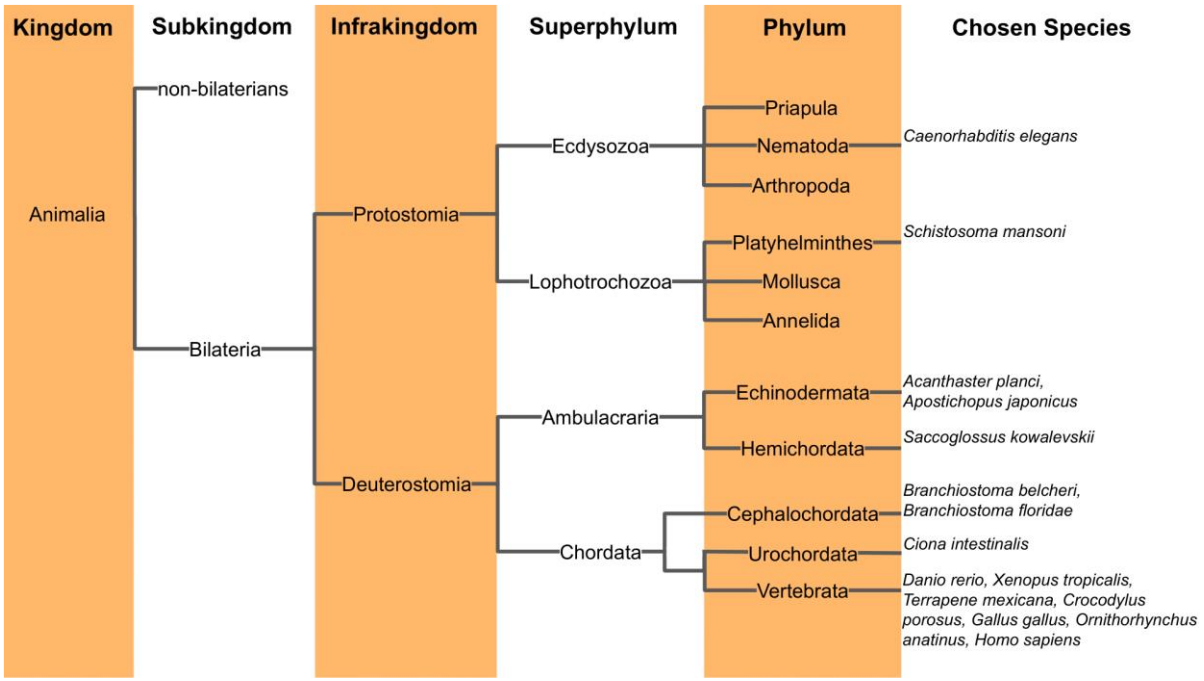


Figure 8.1 - Phylogenic relationship of bilaterians. Taxonomic categories are indicated at the top of the figure and the chosen species on the right. Chordata ramification events were recently reassessed. Accumulating evidence suggests that cephalochordates diverged first, with tunicates and vertebrates forming a sister group.
Source: Adapted from SATOH *et al*. [123]

Interestingly, the protostomes *Schistosoma mansoni* and *Caenorhabditis elegans* show a sharp frequency peak near 100 bp similar to the one observed in the Olfactores species (Urochordata and Vertebrata). Basal groups of Deuterostomia, however, display a single peak around 10^3 bp and lack the sharp minimal intron peak present in other groups (Figure 8.2). Indeed, a minimal intron sharp peak can neither be observed in Cephalochordata and Ambulacraria species when the intron distribution is displayed on a not logarithmic scale (Figure 8.3).

The fact that a narrow characteristic peak of minimal introns is observed in protostomes [34,77] and it is widespread within eukaryotes argues in favour of the hypothesis that the last common ancestor between protostomes and deuterostomes displayed minimal introns. [34,77] However, representatives from Echinodermata, Hemichordata and Cephalochordata do not show a minimal introns peak (Figure 8.1). Furthermore, these three deuterostomes phylum do not represent a monophyletic

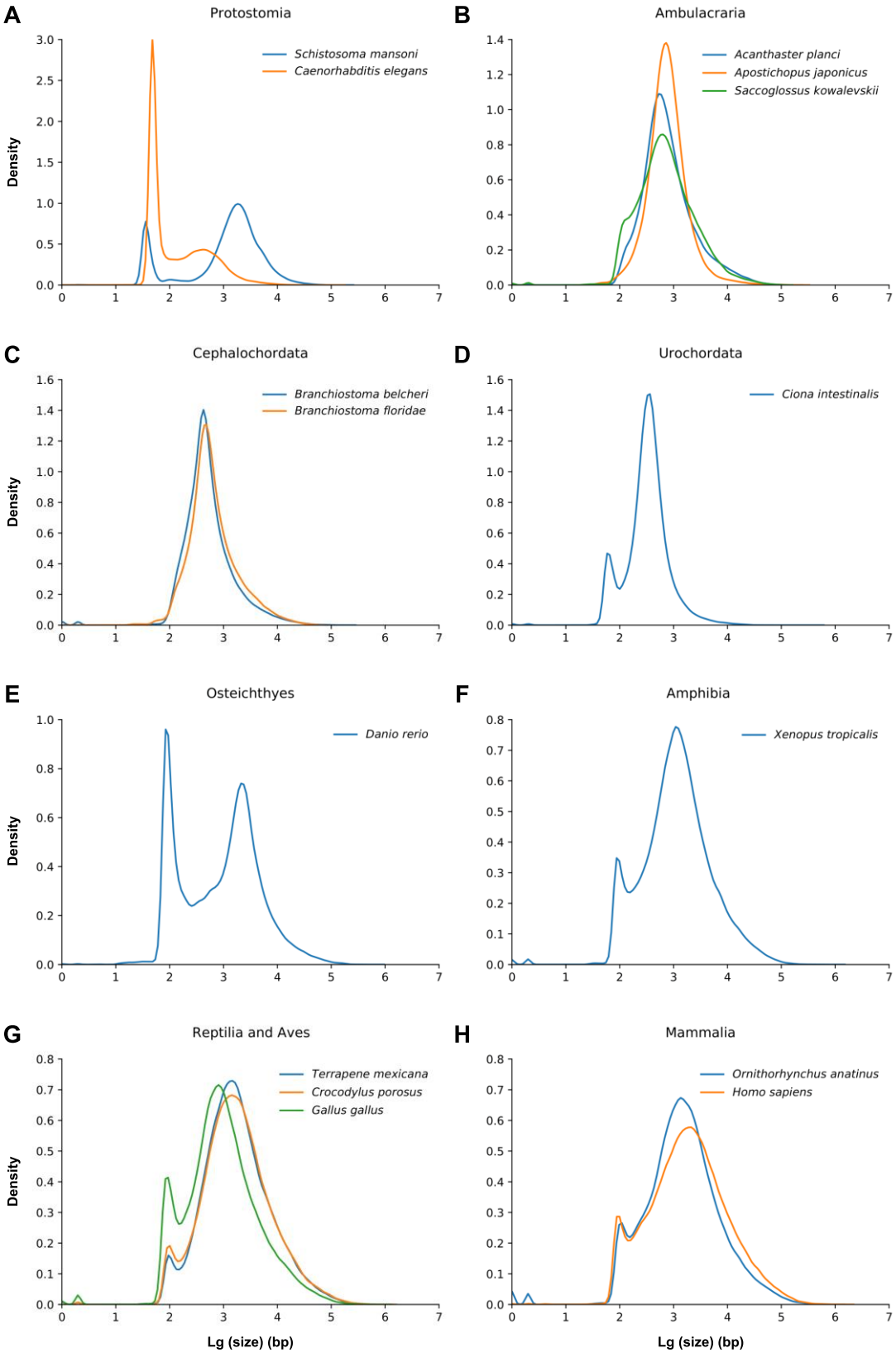


Figure 8.2 - The intron size evolution of bilaterians. Species from protostomes (a), Ambulacraria (b), Cephalochordata (c), Urochordata (d) and different classes of vertebrates (e-h) are shown. Ambulacraria and Cephalochordata representatives differ from other groups by the absence of a minimal intron peak. Intron size distribution was calculated using Kernel Density Estimation (KDE) method and the intron size is shown on a logarithmic scale of base 10.

Source: By the author.

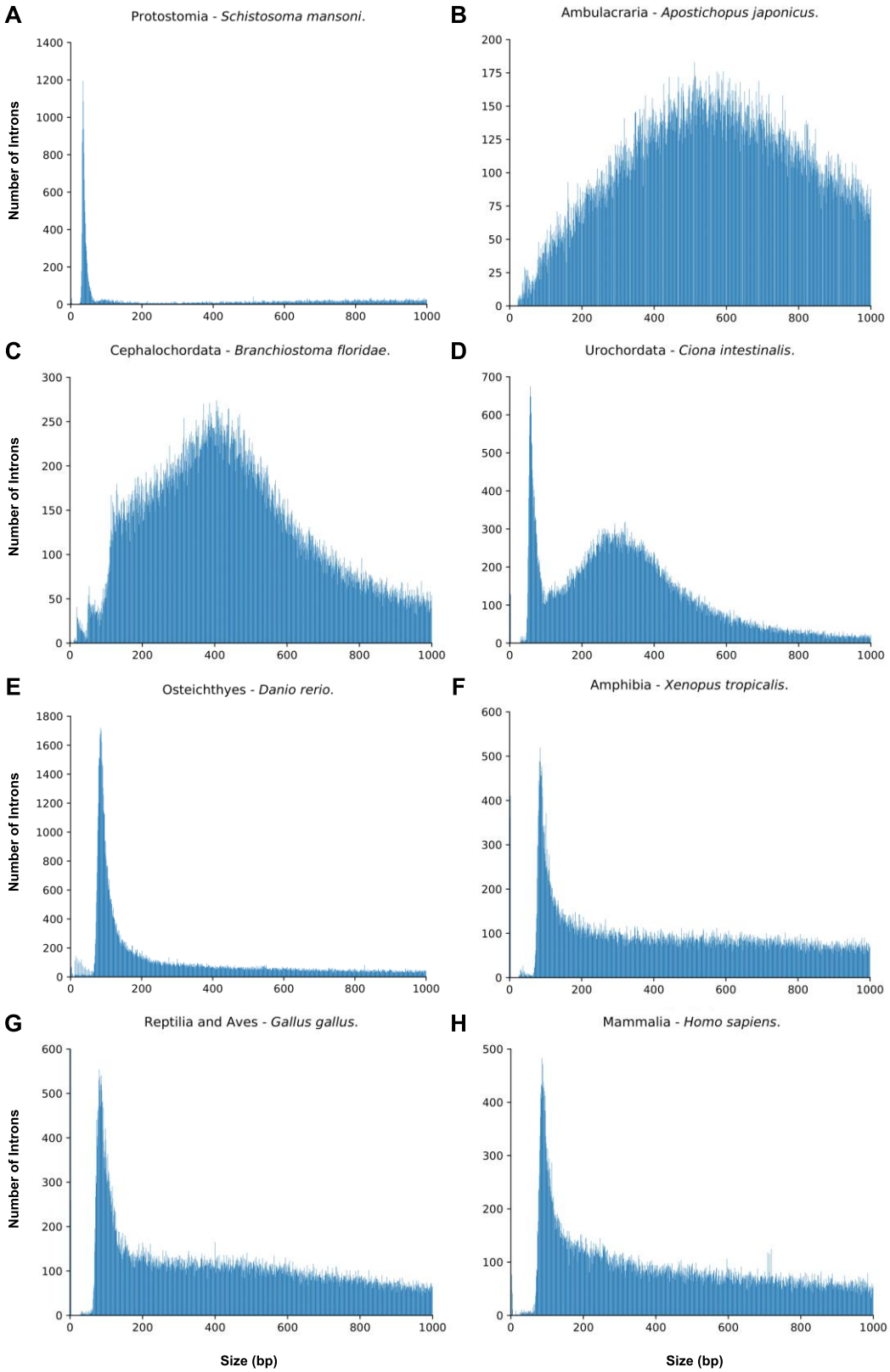


Figure 8.3 - Intron size distribution of bilaterians representatives. Histogram of the number of introns per size is represented for a chosen species of Protostomia (a), Ambulacraria (b), Cephalochordata (c), Urochordata (d) and different classes of Vertebrata (e-h). Ambulacraria and Cephalochordata groups do not show the typical sharp peak of minimal introns. Urochordata species shows a minimal intron peak followed by a wider peak of longer introns. Graphs were made using a bin size of 1 bp and x-axis are limited at 1,000bp.

Source: By the author.

group and therefore, it is not possible to explain such absence by a single event of loss. It should be noted that we do not propose that such an event of loss represent the actual loss of the minimal introns, but rather an event that modifies the splicing machinery and removes the evolutionary pressure to keep part of the population of introns at low sizes. After such an event, minimal introns sizes would be allowed to drift and eventually acquire a similar profile to that observed for regular-sized introns. Considering that scenario, we envisage two possible explanations for the observed distributions. The first explanation is that two independent events of loss occurred, one at the base of Ambulacraria and another at the base of Cephalochordata. The second explanation would involve an event of loss occurring at the base of Deuterostomia followed by an event of re-establishment of the machinery responsible for the evolutionary pressure for minimal introns at the base of Olfactores (Urochordata + Vertebrata).

Curiously, Echinodermata, Hemichordata and Cephalochordata organisms show a regular intron peak size of one order of magnitude smaller to what is observed in the Vertebrata (Figure 8.2.) It is therefore tempting to assume that changes in regular introns sizes might be somehow connected to the presence of minimal introns. However, Urochordata representative shows a typical minimal intron sharp peak followed by a regular intron peak with a size distribution that is similar to what is observed in Cephalochordata, Echinodermata and Hemichordata (Figure 8.3d). Therefore, there is still no convincing link between the presence of minimal introns and the size distribution of regular-sized introns.

8.2 Minimal intron peak is conserved in the Vertebrata phylum

Results from the intron size distribution show a slight shift of minimal intron peak towards higher values in the evolution of the Animalia kingdom, where Vertebrata group displays a peak with the largest size of minimal introns (Figure 8.4a). Whereas minimal intron modal value is 34 bp, 47 bp, and 57 bp in the *Schistosoma mansoni*, *Caenorhabditis elegans*, and *Ciona intestinalis* species, respectively, this number fluctuates around 85 bp in the Vertebrata phylum. Furthermore, minimal intron peak

location seems to be conserved from lower vertebrates to mammals (Figure 8.4b). Indeed, there is little variation of the modal size value and minimal intron range, according to our definition (Table 8.1).

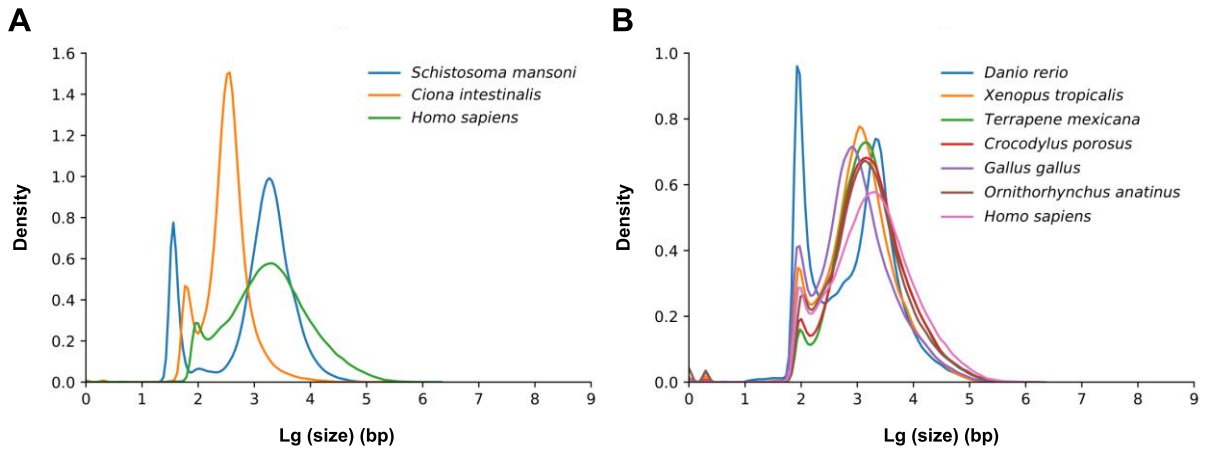


Figure 8.4 - Intron size distribution of Vertebrata. In relation to other bilaterian animals, vertebrates show a minimal intron peak with a slight shift to the right (a). The minimal intron peak localization, however, is conserved from lower vertebrates to mammals (b). Graphs were created using the Kernel Density Estimation (KDE) method.
Source: By the author.

Table 8.1 - Minimal introns from Vertebrata.

Species	Modal Size (bp)	Minimal Intron Range (bp)
Homo sapiens	85	78-92
Ornithorhynchus anatinus	91	81-101
Gallus gallus	80	71-89
Crocodylus porosus	89	79-99
Terrapene mexicana triunguis	90	79-101
Xenopus tropicalis	84	77-91
Danio rerio	84	76-92

Source: By the author.

8.3 Materials and methods

8.3.1 Extraction of intron information from the chosen genomes

Intron information was extracted from annotation genomes files of GFF type via scripts written in Python programming language (version 3.7.3) with the support of Pandas library (version 0.23.4) for data manipulation improvement. Information regarding the version of genomes used from different Animalia species is summarized in Table 8.2. Intronic coordinates in the genome were inferred based on exonic coordinates. In the end, a dataset for each species was built with information regarding the parent chromosome, gene and mRNA of the intron, plus its position in the gene, coordinates, size and strandedness.

8.3.2 Intron size distribution and minimal intron definition

After calculating intron size based on its coordinates, values were plotted using the non-parametric Kernel Density Estimation method (KDE) to estimate the probability density function of intron size. Histogram of absolute counting values for each size was also plotted using a bin size of 1 bp. If the species presents a narrow peak near the minimum size, the minimal introns range were determined based on the following criteria. The low limit is defined as the minimum intron size with at least half of the number of introns at the modal value. The low limit definition step aims to remove very small introns that may represent annotation artefacts of the genome. The upper limit is defined by mirroring the low limit distance to the peak. For example, if the intron modal value is at 40 bp and its low limit is at 35 bp, the upper limit would be 45 bp. Regular-sized introns were defined as those larger than the upper limit. The analysis was conducted using Python programming language (version 3.7.3) with support of Pandas (version 0.23.4), Numpy (version 1.16.4), Seaborn (version 0.9.0) and Matplotlib (version 2.2.3) libraries.

Table 8.2 - Genome version of analysed Animalia species.

Species	Genome Version	Accession Number [a]
Homo sapiens	GRCh37.p13	GCA_000001405.14
Ornithorhynchus anatinus	mOrnAna1.p.v1	GCF_004115215.1
Gallus gallus	GRCg6a	GCF_000002315.6
Crocodylus porosus	CroPor_comp1	GCF_001723895.1
Terrapene mexicana triunguis	T_m_triunguis-2.0	GCF_002925995.2
Xenopus tropicalis	Xenopus_tropicalis_v9.1	GCF_000004195.3
Danio rerio	GRCz11	GCA_000002035.4
Ciona intestinalis	KH	GCF_000224145.3
Branchiostoma belcheri	Haploidv18h27	GCF_001625305.1
Branchiostoma floridae	Version 2	GCF_000003815.1
Apostichopus japonicus	ASM275485v1	GCA_002754855.1
Acanthaster planci	OKI-Apl_1.0	GCF_001949145.1
Saccoglossus kowalevskii	Skow_1.1	GCF_000003605.2
Schistosoma mansoni [b]	v7.1	-
Caenorhabditis elegans	WBcel235	GCF_000002985.6

[a] From NCBI database.
[b] Downloaded from FTP link: ftp://ftp.sanger.ac.uk/pub/project/pathogens/Schistosoma/mansoni/v7/.
Source: By the author.

Chapter 9

THE EFFECT OF HOMEOTHERMY IN THE EVOLUTION OF MINIMAL INTRONS

The homeothermy endothermy feature observed in Aves and Mammalia vertebrate classes might be one of the most important events during vertebrate evolution. Although many other organisms are able to sustain a body temperature higher than that of the environment, such as the leatherback sea turtle, [129] female pythons [130] and large-size tunas, [131] the endothermy observed in those groups are not as outstanding as the one of birds and mammals. The source and magnitude of heat production, the stable and markedly elevated body temperature of birds and mammals are considered unique among the Metazoa. The heating strategy adopted by these animals is a result of a combination of high resting, the presence of aerobically heat production in virtually all soft tissues, and a good insulation system able to significantly decrease heat loss. [132] Probably as a consequence of high body temperature, the genome of warm-blooded vertebrates differs from the poikilotherms animals in many aspects, such as GC content increase; [133] genome fragmentation into DNA segments of homogeneous GC% composition known as isochores; [134] and intron lengthening. [135] Interestingly, minimal introns of the Vertebrata group do not show an apparent lengthening, but instead a highly conserved peak size from lower vertebrates to mammals. In this chapter, we explored the changes of minimal introns during the vertebrate evolution that may be related to the increase in the body temperature of those animals.

9.1 Increased body temperature segregates minimal introns into two populations

To verify the differential effect of temperature on introns during the evolution of vertebrates and how the acquisition of homeothermy feature may have affected them, we analysed the GC content of two introns populations: the minimal introns and long introns. In this analysis, long introns, defined as those longer than 1,000 bp, were used instead of our previously defined regular-sized introns in order to assure a population with very distinct characteristics from the minimal introns (Figure 9.1). Curiously, the distribution of minimal introns, in regard to GC content, displays two peaks that allow the segregation into two subpopulations of low and high GC content on Reptilia, Aves and Mammalia classes (Figure 9.2a). Long introns, however, always show a unimodal distribution, where the GC content of the peak resembles the GC% of the genome of the species (Figure 9.2b).

Analysis of the GC content of minimal introns in several vertebrate species suggests a gradual transition from lower to higher vertebrates. The fish representative *Danio rerio* shows a unimodal distribution at low GC content. Amphibian still presents a single peak, but with asymmetrical distribution where a longer tail towards higher GC% can be observed. Minimal introns of reptiles can already be divided into two subpopulations, where the low GC% subpopulation still predominates. Birds and mammals display a prevalence of high GC% subpopulation (Figure 9.2a).

It has been previously described that an increase in body temperature leads to higher GC content and to the observation of a mosaic distribution into homogeneous GC composition regions known as isochores in the vertebrate genome. [133-134] Human genome, for example, can be divided into isochores partitioned in four different families. Light isochores represent 63% of the genome and have a mean GC content of ~39%. Three subpopulations of heavy isochores represent 24%, 7.5% and 4.7% of the genome and have a mean GC content of 44%, 47% and 52%, respectively. [136] Therefore, it is probable that the same underlying phenomenon that leads to a general increase in GC in the genome due to the increase of body temperature is also increasing the GC content of minimal introns. However, the observed shift of GC content in minimal introns is remarkable as the majority of minimal introns underwent a large increase in GC content with a peak of ~70%.

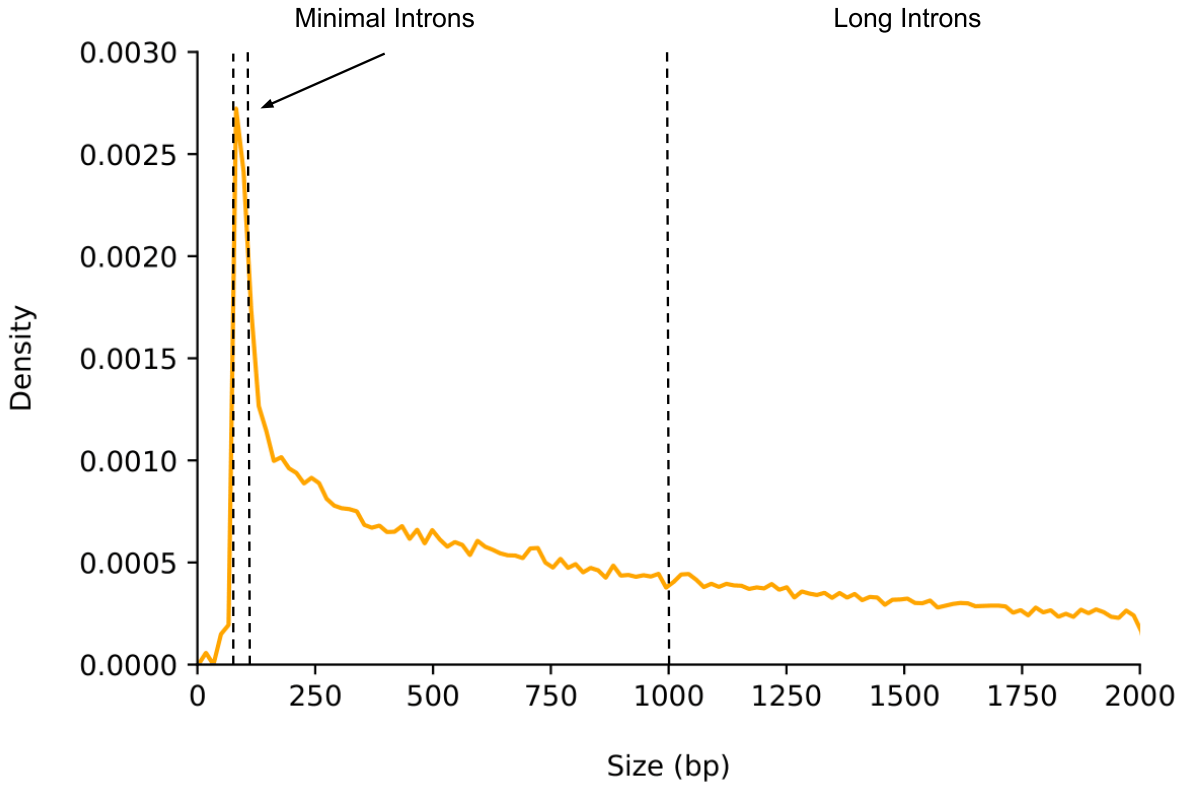


Figure 9.1 - Size distribution of human introns. Introns size distribution shows a prevalence of a minimum intron size peak. This peak is species-specific and it varies from 80-90 bp in vertebrates. Minimal introns are defined as those near the peak, whereas long introns are defined as those longer than 1,000 bp. The x-axis was truncated on 2,000 bp length, however, it extends until 1 Mbp without any further peak in humans.

Source: By the author.

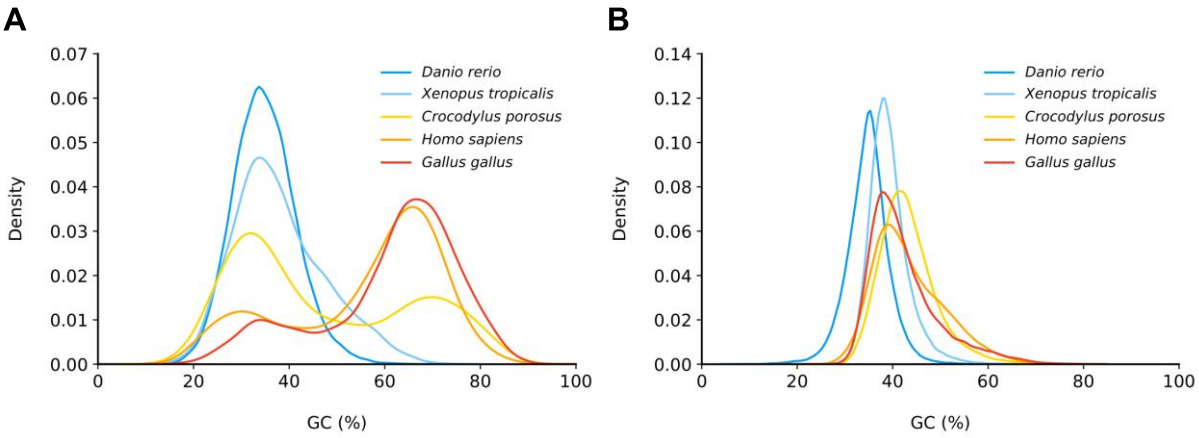


Figure 9.2 - GC content distribution of introns from vertebrates. Except for the fish and frog representatives, minimal introns can be divided into two defined populations of low and high GC% with peaks around 30% and 70%, respectively (a). The distribution pattern of minimal introns is suggested to be influenced by the body temperature of the species. Long introns, on the other hand, do not show the same bimodal distribution, but a single GC% peak near the overall GC content of the genome (b). Colours refer to the body temperature of the respective species, from lower temperature (blue) to higher temperature (red).

Source: By the author.

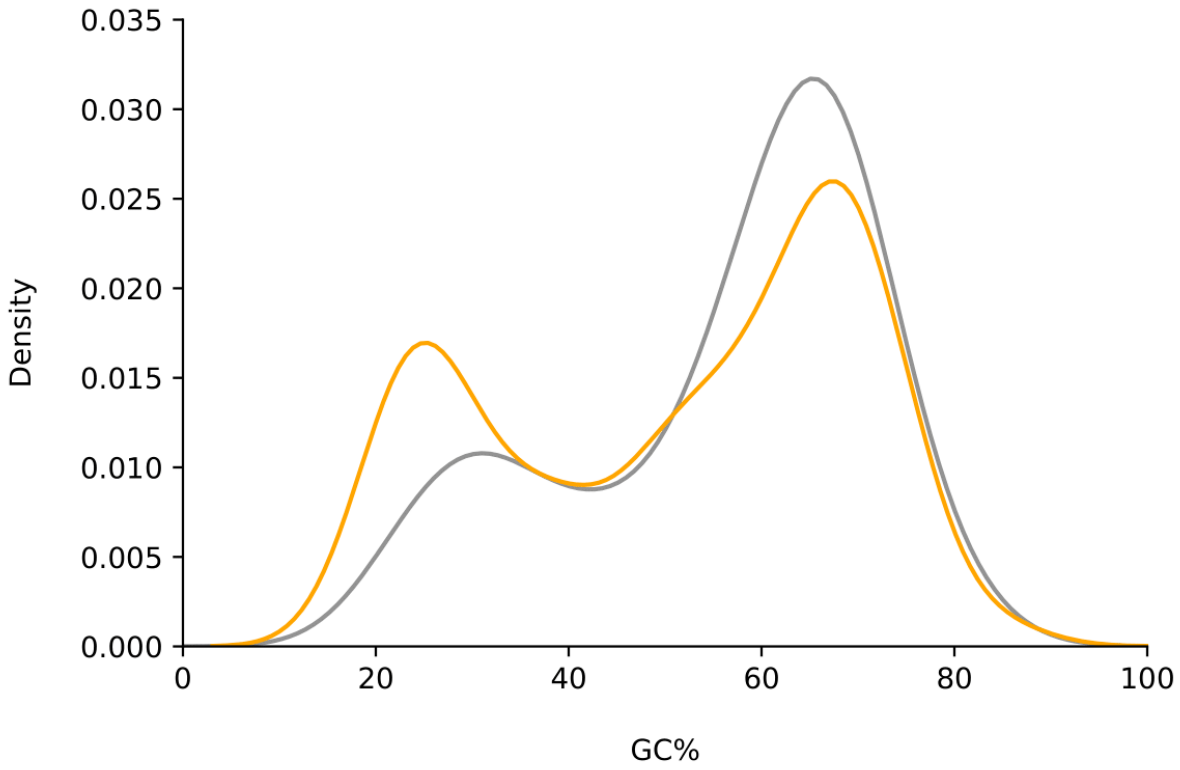


Figure 9.3 - Distribution of GC% of minimal introns expressed in the human testis and other parts of the human body. Human minimal introns are divided into low and high GC% populations if lower or greater than a GC content of 43.98%. Minimal introns of high GC% are predominant in the human body (shown in grey). However, minimal introns from genes of which expression is higher in the human testis (shown in orange), where the temperature is 2-4 ºC below body temperature, [137] shows an increment of the proportion of the population of introns with low GC%. Only minimal introns from genes with tissue score greater than 5 were analysed in the testis.

Source: By the author.

In order to understand if the temperature is a determinant factor behind the observed increase in the GC content of minimal introns, we decided to verify if tissues of the same organism that displayed different temperatures have different distributions of the GC content of the minimal introns that reside in genes expressed in those analysed tissues. This approach has the advantage of comparing genes of the same organism, thus minimizing other possible factors inherent to each species that could explain the observed shifts. It has been described that the temperature of the testis is 2-4 ºC lower than the rest of the body [137] and therefore it is possible that the main genes expressed in this tissue might have a different profile regarding the GC content of its minimal introns. Interestingly, genes that were at least five times more expressed in testis than in any other human tissue appear to have a greater proportion of the low

GC% subpopulation of minimal introns when compared to the minimal introns from the other genes (Figure 9.3). To verify if the difference in the distribution of the populations was statistically significant, we divided the introns in two populations of high and low GC%, using the value of 43.98% to separate those populations. We then compared the proportions of minimal introns with low and high GC% in the two populations of genes with a Chi-square test of independence of variables and verified that the differences are indeed significant (p-value of 1.44e-3, Table 9.1). This result suggests that the increase in GC content of minimal introns is related to the expression or processing of the mRNA since the response appears to be related to the temperature of the tissue in which expression occurred. In this way, this indicates that the GC% of minimal introns may be related to transcription, processing or even translation of the gene.

Table 9.1 - Number of minimal introns analysed in the testis and other human tissues.

Tissue	Low GC% Minimal Introns [a]	High GC% Minimal Introns [a]
Testis	55	99
Other	1,427	4,471

[a] Values used in the Chi-square test of independence of variables to verify differences in the proportion of low and high GC% minimal introns from the testis and other human tissues. P-value = 1.44e-3.
Source: By the author.

9.2 Minimal introns as proxies for studying temperature-responsive genes

Our previous analysis showed that minimal introns display a much larger GC% variation in different classes of vertebrates than do longer introns (Figure 9.2). This suggests that the breadth of GC content variation is dependent on the size of the intron. Indeed, a more detailed analysis observing the distribution of introns populations of different lengths confirms a direct dependence of GC content breadth and intron length (Figure 9.4). It is also possible to note that the intermediate GC% value of around 40% is favoured in long introns, but such value tends to be avoided as intron size decreases.

Considering that we verified in the previous section that variation of GC% in introns appears to be linked to their transcription or splicing, we hypothesize that the variation of GC content occurs in functional sequences linked to the regulation of those processes. In this scenario, the wider GC% variation breadth in minimal introns would be a result of the high concentration of functional sequences and as introns get larger these functional sequences would be diluted in non-functional sequences that would be non-responsive to temperature changes.

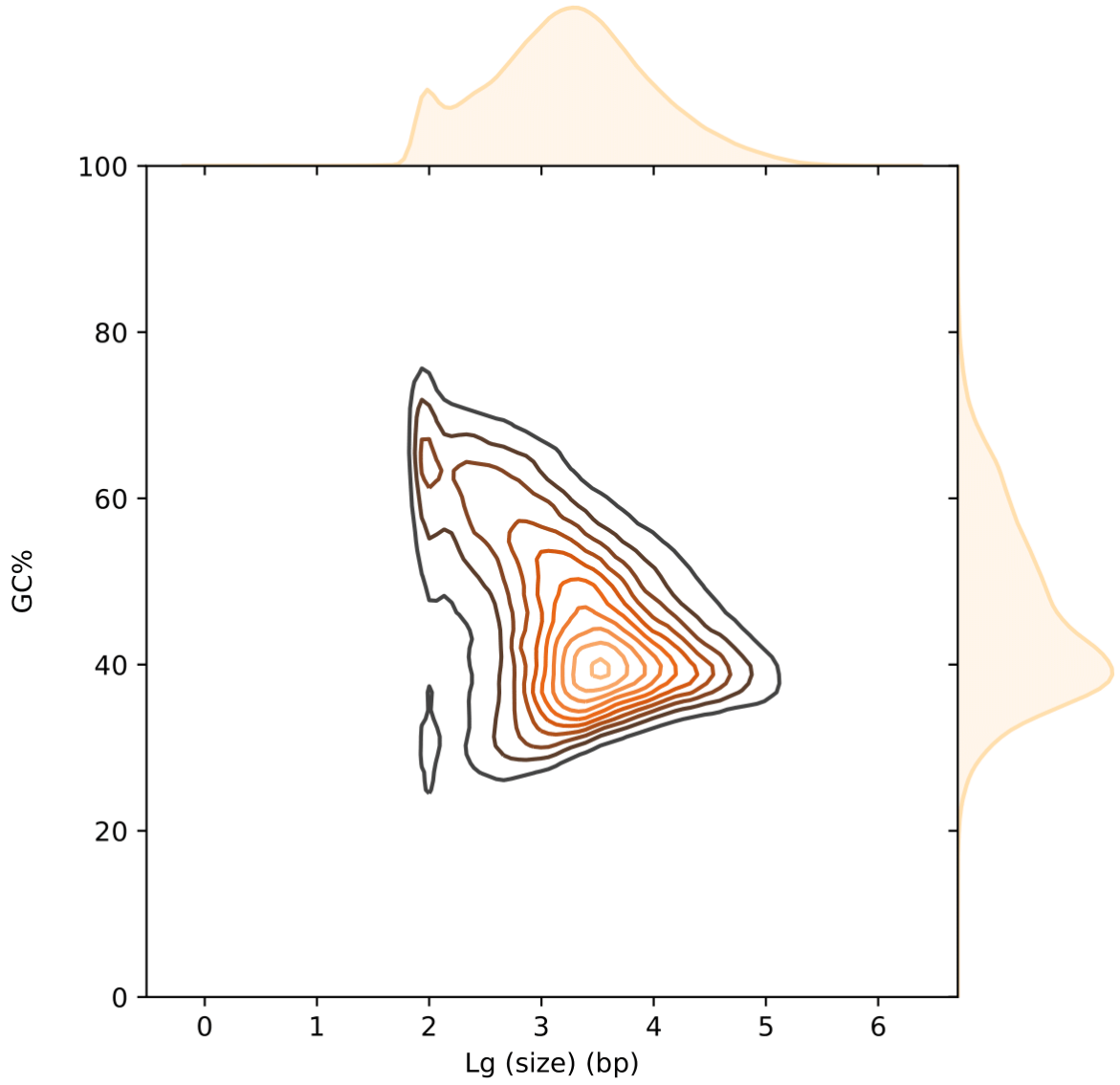


Figure 9.4 - Distribution of GC content of human introns in relation to its size. A two-dimensional KDE plot was used to simultaneously show the GC% distribution of all human nuclear introns derived from protein-coding genes in relation to the distribution of the size of the introns. Unidimensional KDE distributions are shown above y and x-axis. It is possible to note that the breadth of GC content variation is dependent on the size of the intron. Intermediate GC% value of around 40% is favoured in long introns, but it is avoided as intron size decreases.

Source: By the author.

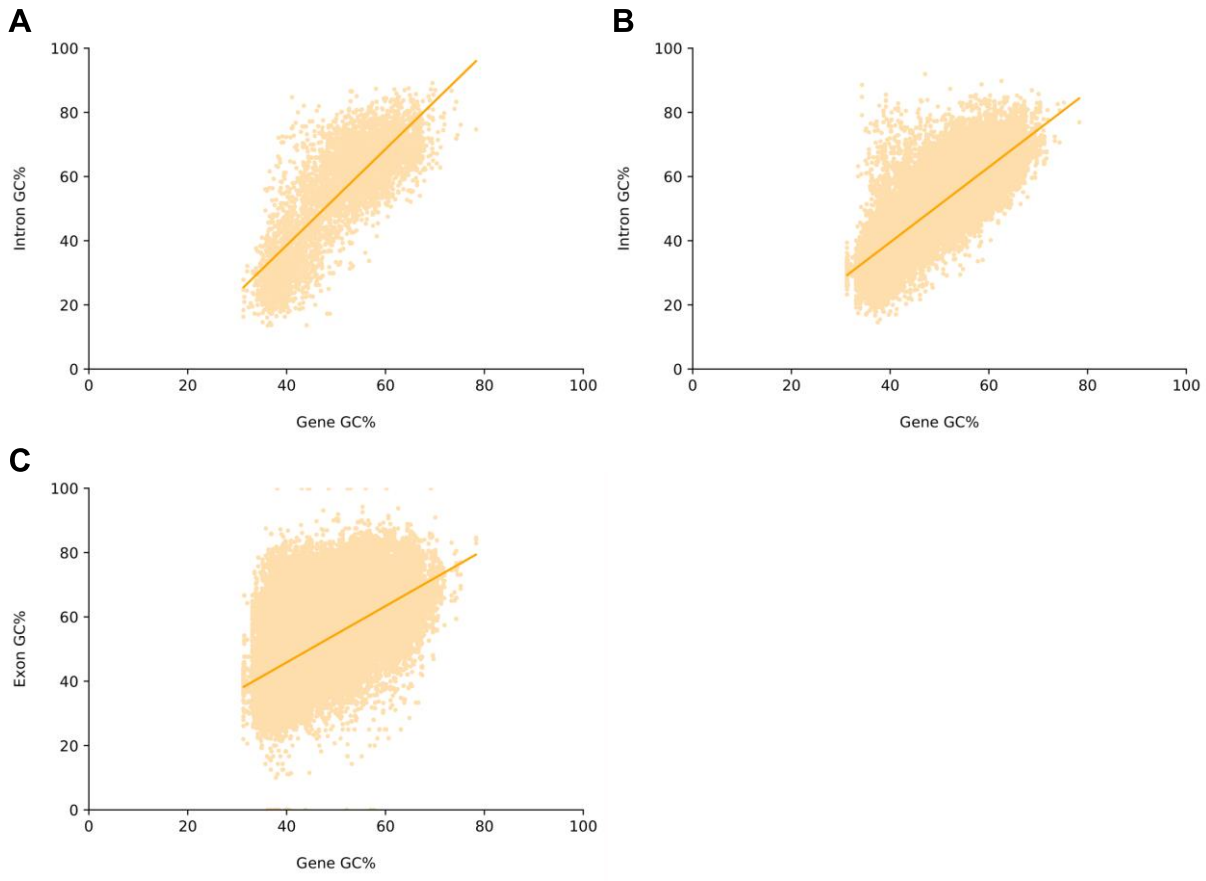


Figure 9.5 - Relationship between GC% of introns/exons and genes. Variation of GC% of minimal introns (intron of 78-92 bp), regular-sized introns (> 92 bp), and exons in relation to the respective gene GC% were modelled using linear regression approach as shown in (a), (b) and (c) respectively. In this analysis, only minimal intron-containing genes were considered, and thus, all analysed exons and introns are derived from minimal intron-containing genes. Calculated models resulted in the coefficient of determination (R^2) and line slope (s) of approximately: 0.67 and 1.50 in (a), 0.73 and 1.17 in (b), 0.47 and 0.88 in (c). Association of intron/exons and genes GC% was also evaluated using the Spearman test that resulted in the correlation coefficients rho (r) of 0.76, 0.87 and 0.70 for (a), (b) and (c) respectively. For linear regression and Spearman tests, all p-values are equal to 0.0.

Source: By the author.

Further evidence for this hypothesis comes from the fact that we find a strong correlation between GC content of minimal introns and the genes harbouring them (Figure 9.5a), verified by the coefficient of determination of linear regression (R^2) of 0.67 and by the correlation coefficients rho (r) from Spearman correlation test of 0.76, both with a p-value of 0.0. The association between minimal introns GC% and the genes GC% is comparable to what was observed between the regular-sized introns GC% and the genes GC% (Figure 9.5b, R^2= 0.73 and r= 0.87, both with a p-value of 0.0), for which we would expect such values since regular-sized introns make up most of the gene sequence. A more modest, but still significant correlation, is observed

between exons GC% and genes GC% with a R^2= 0.47 and a r= 0.70, both with a p-value of 0.0 (Figure 9.5c). This correlation between exons and genes GC% indicates that even an element with several sequence constraints, due to its protein-coding role, has a sequence that is responding to temperature changes.

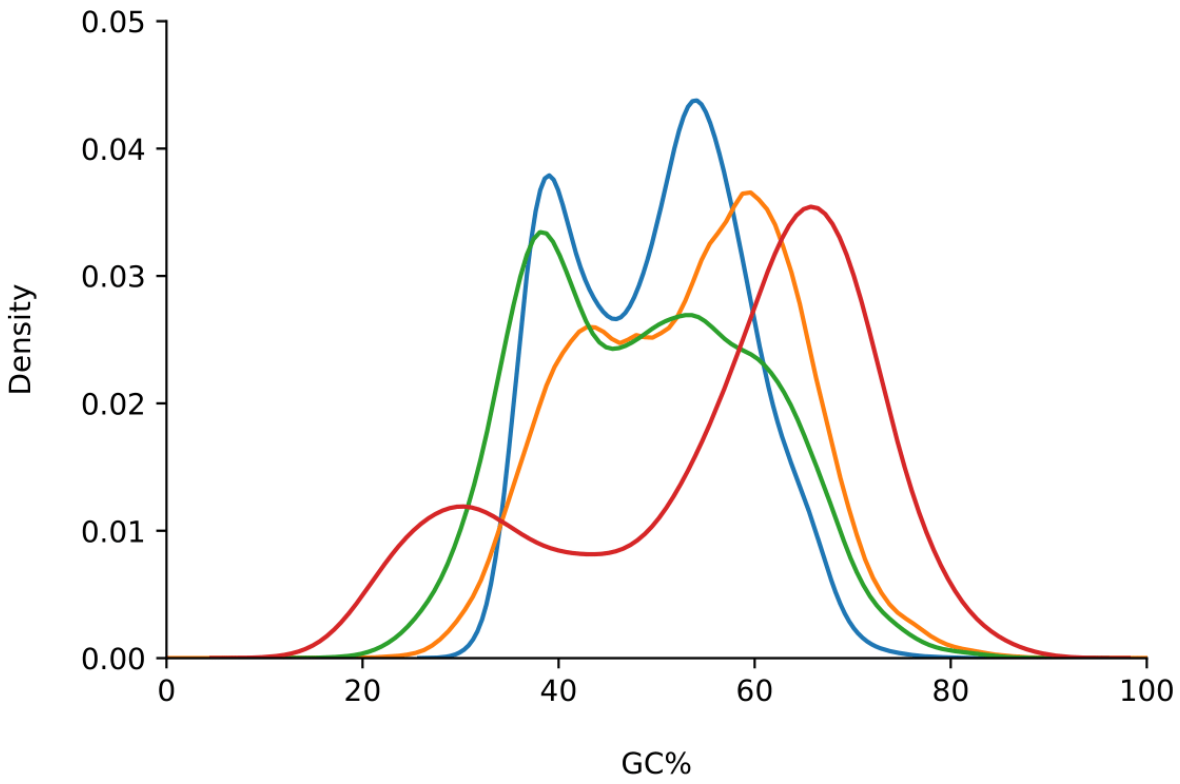


Figure 9.6 - Distribution of GC% of minimal intron-containing genes and its genetic elements. Genes are shown in blue, exons in orange, regular-sized introns in green and minimal introns in red. Introns and exons derived from minimal intron-containing genes show a bimodal distribution of the GC content, as seen in the whole gene itself. Clear segregation of low and high GC% populations is only possible when analysing minimal introns GC content. Probability density functions were estimated using the KDE method.
Source: By the author.

This indicates that changes of GC content in response to an increase in body temperature occur in the gene as a whole, instead of being a specific response of minimal introns. However, analysis of lines slopes (s) of linear regressions from Figure 9.5 shows that a much higher value is obtained for minimal introns (s= 1.5), than that obtained for regular-sized introns (s= 1.17) and exons (s= 0.88). This indicates that the variation of GC content is more extreme in minimal introns than in the other elements of the gene. Indeed, although analysis of the distribution of GC content of these other elements also suggests a bimodal distribution (Figure 9.6), peaks of minimal introns are much more spaced, allowing a clear definition of populations of low GC% and high

GC%. For the other elements and the whole gene, on the other hand, peaks are overlapped and a clear separation of populations is not possible.

Therefore, we argue that, in order to separate these two populations of genes, the use of GC content of the minimal introns as a proxy for the whole gene is more adequate than the direct use of the GC content of the gene itself, since it will allow a good resolution between the two populations due to extreme GC contents while maintaining consistency due to the high correlation observed. This approach would permit the segregation of genes that are responsive to the increase in body temperature with an increased GC content from those that maintained GC content near the ancestral state.

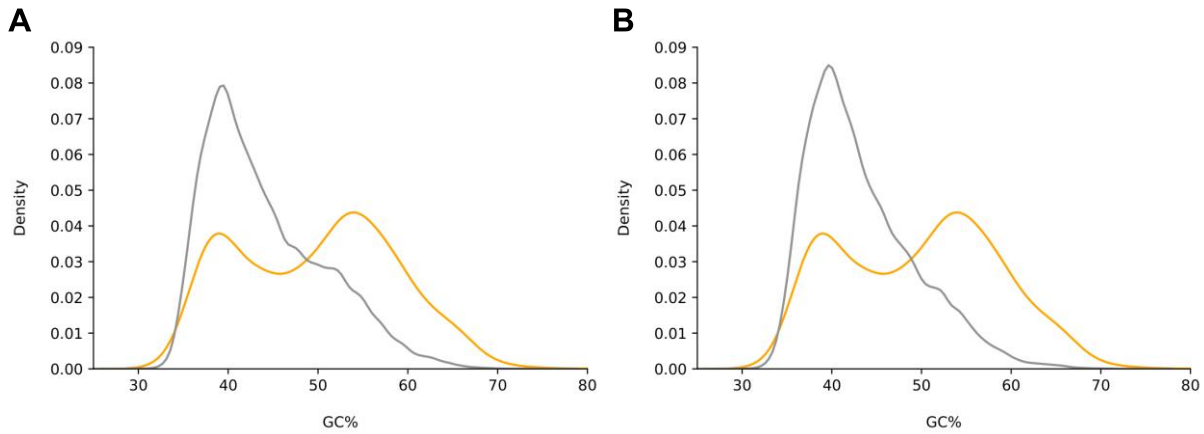


Figure 9.7 - Distribution of GC content of genes with or without minimal introns. Genes that only harbour introns whose size is greater than 92 bp or 200 bp were classified as “genes without minimal introns” shown in grey in (a) and (b) respectively. For both analyses, were classified as “genes with minimal introns”, shown in orange, those genes that contain at least one intron with size between 78 bp to 92 bp. Probability density functions were estimated using the KDE method.
Source: By the author.

Interestingly, if we analyse the distribution of GC content in genes that do not contain minimal introns, we observe a single peak with a small shoulder in higher GC contents (Figure 9.7a). However, if we remove genes containing introns shorter than 200 bp such a shoulder practically disappears (Figure 9.7b), meaning that it represents mostly genes that contain introns near the minimal intron range. This indicates that genes that do not contain minimal introns are mostly non-responsive to temperature changes concerning GC content with few exceptions. One could argue that those genes would be similar in nature to minimal intron with low GC% since both appear to

be non-responsive to increases in body temperature. We think that is not the case since the migration to high GC content in minimal introns is a dynamic process throughout vertebrate evolution (Figure 9.2a). Therefore, we think that low GC% minimal intron-containing genes are subjected to the same evolutive pressure that led to the migration of many minimal intron-containing genes to high GC%, but another evolutive force may counter such pressure. As body temperature increases, the pressure to assume higher GC% would increase and in a larger proportion of genes, this would prevail on the opposing pressure to stay at low GC%. On the other hand, genes without minimal introns would be truly non-responsive to temperature increase concerning GC content since a negligible fraction of them show high GC%.

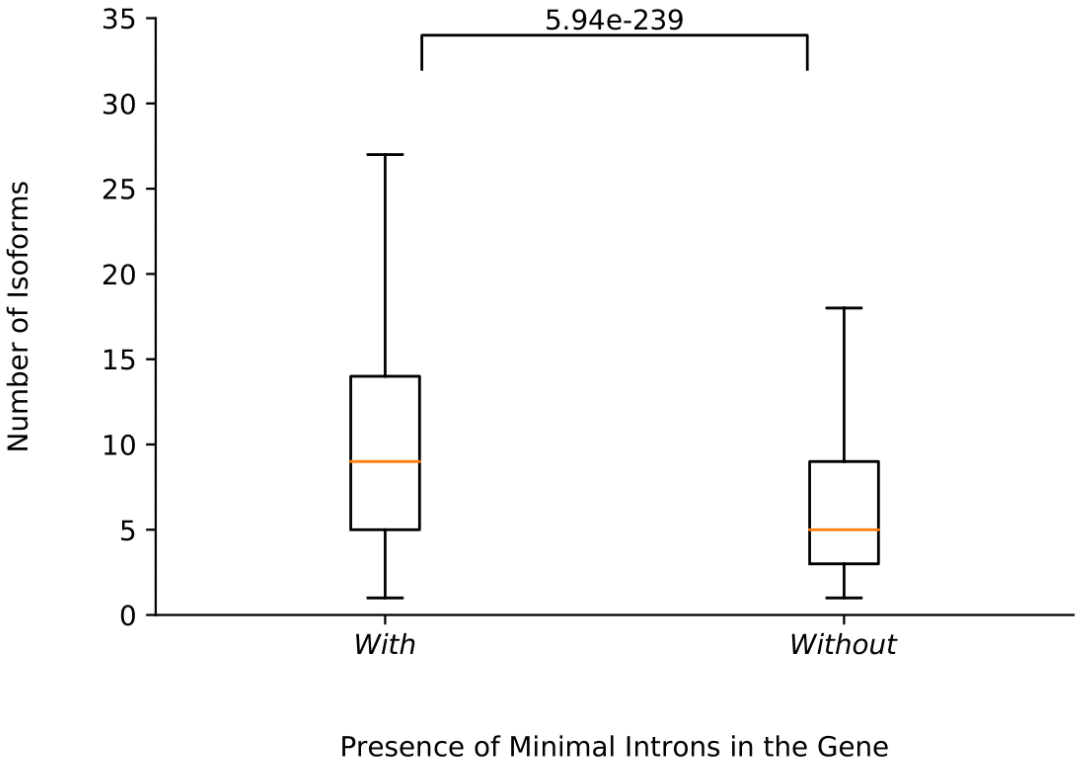


Figure 9.8 - Number of isoforms from the two types of human genes. Minimal intron containing-genes shows a greater proportion of isoforms than genes without minimal introns. The distribution of the number of isoforms in each type of gene is represented by a boxplot. The coloured central line represents the median, the first quartile (25th percentile) by the bottom line of the box, the third quartile (75th percentile) by the upper line of the box, and the maximum and minimum values by the lines at the end of the whiskers. Outliers, if present, are represented by points. P-value from the non-parametric Mann–Whitney U test is shown above the horizontal bracket.

Source: By the author.

Furthermore, while investigating the two types of human genes, we noticed that minimal intron-containing genes show a greater number of isoforms than genes without minimal introns (Figure 9.8). As alternative forms of splicing are one of the main mechanisms involved with isoforms creation, we suggest that transcripts derived from minimal intron-containing genes may have differences in splicing regulation from the genes with no minimal introns.

9.3 Bimodal distribution of minimal introns in extreme GC contents is associated with intron retention levels

Our previous analysis suggests a link between transcription/RNA processing and GC content variation, with distinct behaviours in terms of GC content variation correlated to elements (intron/exons) of the transcript. It has been previously described that the processing of mRNA is directly affected by changes in temperature, with an increase of IR events observed when cells are submitted to a heat shock. [71] Therefore, it is reasonable to suppose that a long-term increase in body temperature might elicit similar processes and trigger evolutionary pressures to avoid a widespread phenomenon of IR. Moreover, it has been described that IR has a regulatory role in gene expression and that occurs more frequently in shorter introns that are spliced via intron definition. [66,138] In this context, it is reasonable to suppose that one of the possible processes affected by the increase in body temperature in vertebrates is the regulation of IR.

In order to obtain further insights into the dynamic of the IR process in minimal introns of different GC contents, we divided them into three populations of low, intermediate and high GC% (Figure 9.9a). Libraries of RNA-Seq data from nuclear immature transcripts (non-polyadenylated transcripts), nuclear mature transcripts (polyadenylated transcripts) and cytoplasmic polyadenylated transcripts were used to calculate the percentage of intron retention (PIR). As expected, we see a progression with nuclear non-polyadenylated transcripts displaying the highest levels of PIR for all populations of minimal introns and cytoplasmic polyadenylated transcripts showing the lowest PIR levels (Figure 9.9b-d). The large difference in PIR levels between nuclear

and cytoplasmic samples may be a result of storage of transcripts with retained introns in the nucleus, a later transcript processing, or even a rapid decay of transcripts with retained introns in the cytoplasm. Interestingly, minimal introns with intermediate GC% have the highest IR fraction of the three minimal introns populations (Figure 9.9b-d).

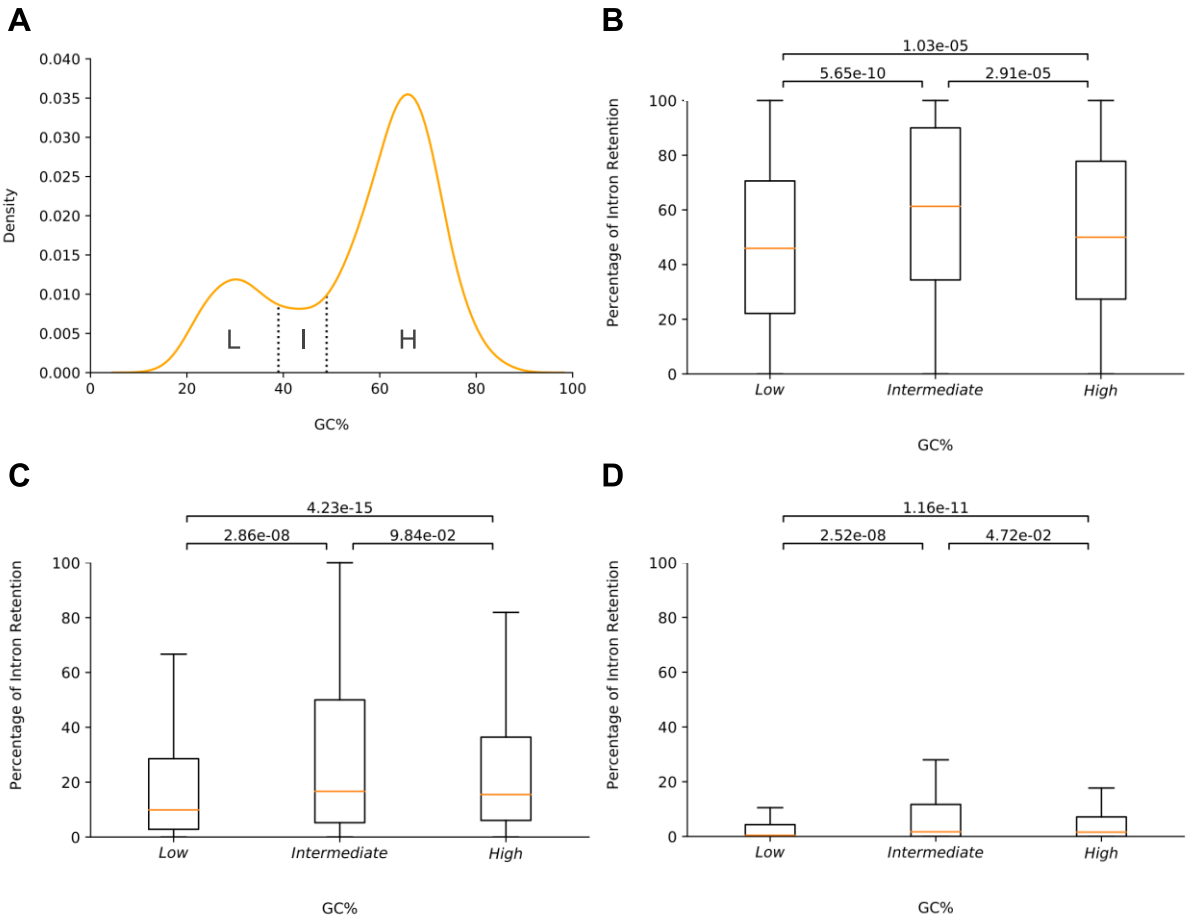


Figure 9.9 - Distribution of percentage of retention from three groups of human minimal introns. For this analysis, minimal introns were divided into three populations based on the GC content: low GC% (<= 43.98-5, L), intermediate GC% (43.98±5, I), and high GC% (>=43.98+5, H) (a). Analysis from RNA-seq of immature transcripts found in the nucleus of B-lymphocyte cells (b), mature (polyadenylated) nuclear transcripts (c) and cytoplasmic polyadenylated transcripts (d) are shown. The distribution of the PIR values from each group of minimal introns is represented as a boxplot. The coloured central line represents the median, the first quartile (25th percentile) by the bottom line of the box, the third quartile (75th percentile) by the upper line of the box, and the maximum and minimum values by the lines at the end of the whiskers. Outliers, if present, are represented by points. P-values from the non-parametric Mann–Whitney U test are shown above the horizontal brackets.

Source: By the author.

Considering that intermediate GC content is avoided by minimal introns, we wonder if such increased levels of IR might be associated with the evolutionary pressure keeping the majority of minimal introns at extreme GC contents. Moreover, if

we assume that an intermediate GC content might be the cause of increased IR, the transition of a minimal intron from low to high GC contents might be cumbersome to introns in which fine regulation of IR is important.

9.4 Transcription abortion does not explain the need for GC% increase

Analysis of IR and gene function of minimal introns have raised some hypotheses that may explain the distribution pattern of minimal introns GC% found in genomes of high body temperature species. However, why in the first place does the GC content have to be increased when in high temperature? A previous study has shown that high GC% genes increase mRNA levels in mammalian cells and suggested that this process may be related to an efficient transcription process. [139] With this in mind, we hypothesize that the transition to high GC% could be associated with an increase in the stability of the RNA-DNA hybrid during the transcription process. The RNA-DNA hybrid of 8 bp is maintained during RNA Polymerase II elongation stage and its disruption may be the pivotal signal that leads to transcription termination. [140] In this way, we speculate that minimal introns with low GC% could have a higher probability of RNA-DNA disruption in high-temperature systems due to its weaker base-pairing formations. This could represent a termination signal that would lead to premature transcription abortion and thus decrease the mRNA level.

To test our hypothesis, we decided to count the reads from RNA-Seq data of non-polyadenylated transcripts found in the nucleus mapping in exons upstream and downstream of human minimal introns. In this analysis, we divided minimal introns in three populations of low, intermediate and high GC%. If our hypothesis is correct, we would expect that count of downstream exons of low GC% minimal introns would be lower than the count number of upstream exons. However, the ratio of upstream and downstream exons counts of the three minimal introns populations did not show any relevant difference (Figure 9.10). As so, this result could not corroborate with our transcription abortion hypothesis. Is still unknown the mechanism behind the need for GC% transition in high-temperature species.

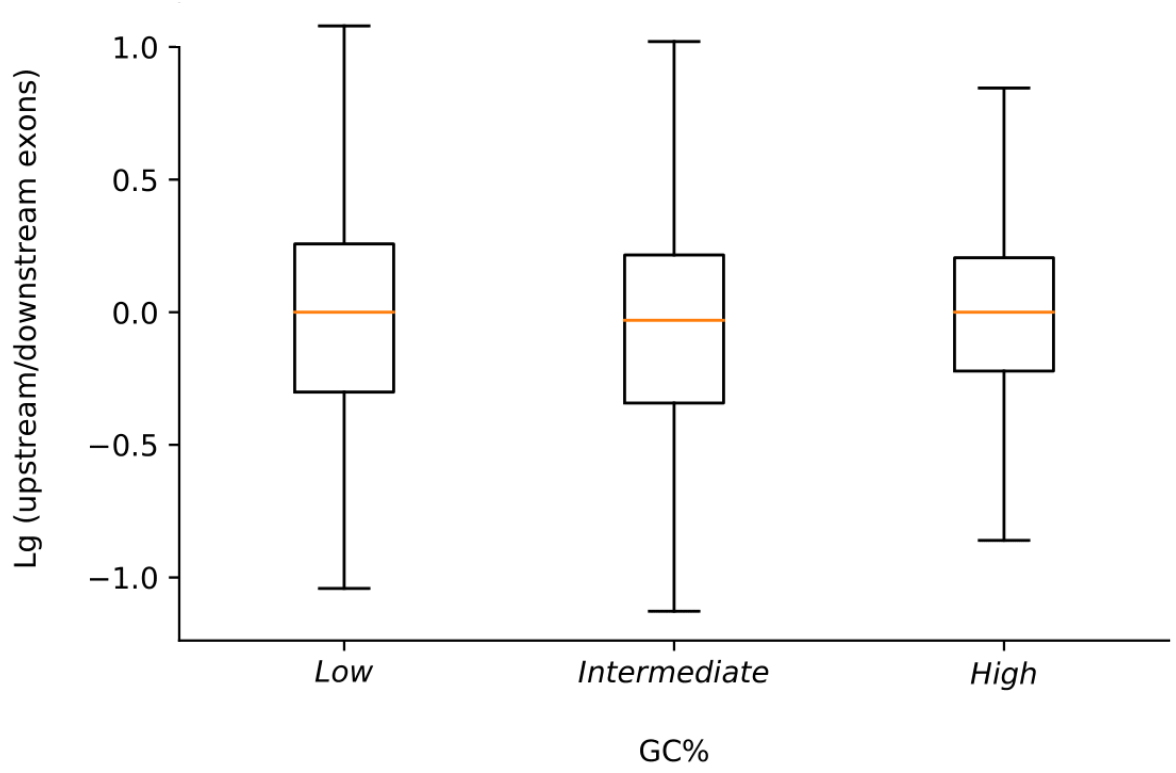


Figure 9.10 - Analysis of upstream and downstream exons of human minimal introns. RNA-Seq data from immature transcripts (non-polyadenylated transcripts) found in the nucleus of B-lymphocyte cells was used to count the reads that mapped to the respective exons. Count of exons adjacent from low GC% (<=43.98-5), intermediate GC% (43.98±5), and high GC% (>=43.98+5) minimal introns are shown. The distribution of the ratio of upstream and downstream exons for each minimal intron group is shown as a boxplot. The coloured central line represents the median, the first quartile (25th percentile) by the bottom line of the box, the third quartile (75th percentile) by the upper line of the box, and the maximum and minimum values by the lines at the end of the whiskers. Outliers, if present, are represented by points. Logarithmic scale of base 10 was used in order to avoid asymmetric values.
Source: By the author.

9.5 Materials and methods

9.5.1 Distribution of introns GC% of vertebrates

Introns from vertebrate species were analysed in terms of their GC content. Introns were divided into minimal introns and long sized introns. Minimal introns were classified by the same criteria used in Chapter 8. In this chapter, we introduced the classification of long introns as those longer than 1,000 bp to sample a population with very distinct characteristics from the minimal introns. To calculate introns GC%, sequences of introns were extracted from genome sequence file using the coordinates

found in genome annotation file as described in Chapter 8. This and the following analysis were conducted in Python programming language (version 3.7.3) with support of Biopython (version 1.72), Pandas (version 0.23.4), Matplotlib (version 2.2.3), and Seaborn (version 0.9.0) libraries unless stated. Intron GC% distribution was estimated using the KDE method.

9.5.2 Intron GC% distribution in the testis and other human tissues

The GC% distribution of minimal introns, estimated by the KDE method, was analysed according to the expression of their respective genes in the testis or other tissues of the human body. Using data from “The Human Protein Atlas”, [141] genes with mRNA levels of, at least, five-fold higher in testis than in all other tissues (tissue score, TS, greater than 5) were defined as testis-enriched genes. In total, 154 minimal introns were classified as belonging to testis-enriched genes. The distribution of the GC content of this population of minimal introns was compared with those from the remaining 5,898 minimal introns. In total, the human genome has 6,052 minimal introns. To obtain the p-value from the Chi-square test of independence of variables, minimal introns were classified as low and high GC% if lower/equal or greater than 43.98%, respectively. The GC% cut-point was defined based on the minimum local point that divided the GC% distribution of all human minimal introns (Figure 9.2a). This analysis was also supported by Numpy (version 1.16.4) and Scipy (version 1.3.1) libraries.

9.5.3 Two-dimensional distribution of introns GC% versus size

Human introns derived from protein-coding genes had their GC% calculated using their genomic coordinates and sequence file as early explained. A two-dimensional KDE distribution chart was used to analyse the GC% distribution of introns entities versus their respective sizes. A one-dimensional KDE distribution is also indicated above y and x-axis. Coloured curves represent different population densities. This analysis was also supported by Numpy (version 1.16.4) library.

9.5.4 Distribution of GC content of introns, exons and human protein-coding genes

As early described, all human protein-coding genes, also the exons and introns derived from genes that harbour at least one minimal intron (introns from 78 bp to 92 bp), have their GC content calculated based on coordinates extracted from genome annotation file and sequence file. Genes were divided into those that contain minimal introns and do not contain minimal introns. Genes without minimal introns were defined as those with only introns greater than 92 bp or 200 bp depending on the analysis. Genes with minimal introns were defined in all analyses as those with at least one minimal intron. Introns derived from minimal intron-containing genes were divided into two populations of minimal introns and regular sized-introns (introns greater than 92 bp). The distributions of GC contents were represented as a probability density function created using the KDE method.

9.5.5 Relationship between GC% of introns/exons and genes

In this analysis, only human minimal intron-containing genes, genes that harbour at least one minimal intron (introns from 78 bp to 92 bp), were considered. The variation of the GC content of minimal introns, regular-sized introns (greater than 92 bp) and exons was analysed in relation to their respective minimal intron-containing gene GC%. Relationships between variables were calculated and plotted using the linear regression model. The Spearman test was also used to evaluate the association between variables. This analysis was also supported by the Scipy (version 1.3.1) library.

9.5.6 Number of isoforms of different types of genes

In this analysis, the human genome used was the GRCh38.p13 version (NCBI accession number GCA_000001405.28) [142] as the Ensembl Biomart tool [143] uses this version of the human genome to calculate the number of isoforms per gene. The new peak of human minimal introns is at 88 bp and its range from 78 bp to 98 bp. Genes were classified in relation to the minimal intron presence and its isoforms number were calculated using the Biomart tool. Statistical analysis was done using the non-

parametric Mann–Whitney U test from Scipy (version 1.3.1) library.

9.5.7 Percentage of IR of low, intermediate and high GC% minimal introns

Data from RNA-seq experiments of non-polyadenylated nuclear transcripts (run number SRR317061), polyadenylated nuclear transcripts (run number SRR315298) and polyadenylated cytoplasmic transcripts (run number SRR307899) of human B-lymphocyte cell line were used to calculate the PIR of minimal introns. Raw RNA-Seq reads can be accessed at the NCBI Gene Expression Omnibus (GEO) [144] under the accession number GSE30567. Data of RNA-seq experiments were downloaded using SRA toolkit (version 2.9.6), trimmed using Trimmomatic software (version 0.39) and mapped to the human genome with Hisat2 software (version 2.1.0). Using the Sequence Alignment/Map (SAM) output file, reads with only one reported alignment to the human genome were selected in order to avoid alignment ambiguity. Using reads and introns coordinates, it was counted as retention if the read mapped to at least five nucleotides from the intron and at least five nucleotides to its adjacent exon, and as splicing, if the read did not match intron coordinates and align to both upstream and downstream exons. Only introns which retention and splicing counts that sums up to five were considered in this analysis. PIR was calculated as the ratio of the retention number by the sum of retention and splicing events times 100. Minimal introns were classified as low GC% (<=43.98-5), intermediate GC% (43.98±5), and high GC% (>=43.98+5). Apart from the already cited libraries, our script that calculated PIR values and plotted the results was also supported by the Scipy (version 1.3.1) statistical library. P-values were obtained from the Mann–Whitney U test.

9.5.8 Counts of upstream and downstream exons of minimal introns

Pre-processed data of RNA-seq experiment of non-polyadenylated nuclear transcripts (run number SRR317061) of PIR analysis was used to count reads that mapped to the upstream and downstream exons of human minimal introns. Based on exons and minimal introns coordinates, exons were classified as upstream or downstream. Minimal introns were classified as low GC% (<=43.98-5), intermediate GC% (43.98±5), and high GC% (>=43.98+5). Later, the logarithm of base 10 (Lg) of the ratio of upstream and downstream exons counts was analysed to both groups of

minimal introns. Logarithmic scale of base 10 was used to avoid an asymmetric parameter. Only reads with one reported alignment to the human genome and that mapped to at least ten nucleotides of the exon were considered in this analysis.

Chapter 10

CELLULAR DIVISION AND MINIMAL INTRONS

Alternative splicing is an important component for the creation of complexity in eukaryotic cells. [60] Among the mechanisms of alternative splicing, IR is one whose relevance has only been understood recently. IR is associated with many human disorders, but also with the regulation of complex cellular processes. In a study involving 16 different types of cancer, it was seen that IR is the alternative form of splicing that most differ between normal and cancerous tissues. Primarily, an increase in IR events was reported to be related to the studied types of cancer. [75] On the other hand, IR is important for regulating the temporal expression of genes during intricate processes such as mammalian spermatogenesis. In this process, it was seen that IR is responsible for safely storing transcripts in polyribosomes until posterior expression days after their synthesis. [72]

These studies, together with our previous analyses on IR events of minimal introns with low, intermediate and high GC%, might give us a clue for better understanding the persistence and importance of this group of introns in high body temperature vertebrates. In this section, we are going to explore how IR, low GC% minimal intron-containing genes expression and division process might be correlated. Firstly, we are going to analyse the functional differences of populations of minimal intron-containing genes and their expression profile within a broad range of human tissues. Followed by a detailed analysis of cancer, as a mitotic system, and spermatogenesis, as a meiotic system.

10.1 Low and high GC% minimal introns are found in genes with different functions

The bimodal distribution of minimal intron-containing genes together with the observation that not only minimal introns were affected by temperature increase, but also the entire genes, raised speculations regarding the nature of genes that have migrated to high GC content and those genes that have continued in the ancestry low GC content. In other words: why have some genes been able to migrate to high GC% and others not?

Table 10.1 - Gene ontology terms enriched in minimal intron-containing genes of low GC% and high GC%.

GENES WITH LOW GC% MINIMAL INTRON			GENES WITH HIGH GC% MINIMAL INTRON		
GO - Biological Process [a]	**Fold Enrichment**	**P-Value**	**GO - Biological Process [a]**	**Fold Enrichment**	**P-Value**
Regulation of mRNA stability	2.98	6.45E-03	Catabolic process	1.29	3.89E-02
Microtubule-based transport	2.88	1.19E-02	Localization	1.16	1.07E-02
Establishment of RNA localization	2.87	1.32E-02	Cellular component organization	1.16	2.64E-02
Cell cycle checkpoint	2.77	2.36E-02	Primary metabolic process	1.16	2.69E-05
Mitotic cell cycle phase	2.64	2.09E-03	Organic substance metabolic process	1.16	9.87E-06
Meiotic cell cycle	2.63	4.11E-02	Cellular metabolic process	1.15	7.84E-05
mRNA splicing, via spliceosome	2.54	3.14E-03	Nitrogen compound metabolic process	1.14	1.03E-02
Nucleocytoplasmic transport	2.49	3.86E-02	-	-	-
Chromosome segregation	2.45	4.49E-02	-	-	-
Negative regulation of cell cycle process	2.37	1.68E-02	-	-	-

[a] The analysis was performed using the GO enrichment program powered by Panther software. [145] Test was conducted using Fisher's exact followed by Bonferroni correction. In the table, only the first ten overrepresented terms are shown. Fold enrichment values between 0 and 1 indicate underrepresented terms.
Source: By the author.

In order to investigate if genes with low and high GC% were different from each other, we decided to perform an analysis to detect enrichment of gene ontology terms. In this analysis, human protein-coding genes with at least one minimal intron were divided according to its minimal intron GC content into low and high GC%. Interestingly, low GC% minimal intron-containing genes were associated with biological processes linked to RNA and DNA molecules, such as mRNA stability, chromosome segregation, cell cycle and many others. High GC% minimal intron-containing genes, however, are enriched with terms mainly related to metabolic processes, which are not directly related to DNA or RNA molecules (Table 10.1).

Taken together, the results of PIR analysis with the results of gene ontology terms might explain why some minimal introns could not migrate to high GC content. We speculate that the effects caused by an increase in IR events due to higher levels of GC% may be too harmful to genes related to essential cellular processes, in a way that their transition to high GC% may be blocked. Genes related to accessories process might also experience an impaired function, but with an extent that does not impede their transition. It is still not known, however, if minimal introns with low GC% cope with the consequences of its GC content in those temperatures or if another solution, rather than changing GC content, was adopted by this type of genes.

10.2 Low GC% minimal intron-containing genes are highly expressed in dividing cells

Considering that genes that harbour minimal intron with low GC% display a tendency to be connected with more basic cellular functions, especially those connected with cellular division and chromosomes, we decided to investigate the distribution of expression of those genes in different human tissues and cells. Firstly, genes with and without minimal introns were analysed based on its tissue specificity measured by the Tau (τ) metric. This metric varies between 0 to 1, where 0 means that the same expression profile of a gene is observed in all analysed tissues/cells, whereas 1 refers to a gene which is expressed in only one of the analysed tissues/cells. Interestingly, genes without minimal introns show greater tissue specificity than those

with minimal introns (Figure 10.1a). This result suggests that genes without minimal introns might contribute to the specialization of certain human tissues/cells. On the other hand, genes with minimal introns display less specificity to certain tissues, this indicates that their expression profile is more distributed in many tissues.

More importantly, minimal intron-containing genes were divided in terms of its minimal introns GC content: low, intermediate and high GC%. As expected, genes with minimal introns of low GC% show the lowest tissue specificity values (Figure 10.1b). This result is consistent with the notion that low GC% minimal intron-containing genes are related to vital biological processes such as cell division. This indicates the basal function of this population of genes. On the other hand, as genes with minimal introns of high GC% are related to metabolism, those genes might be more related to the specialization of certain human tissues/cells than to basal cellular functions.

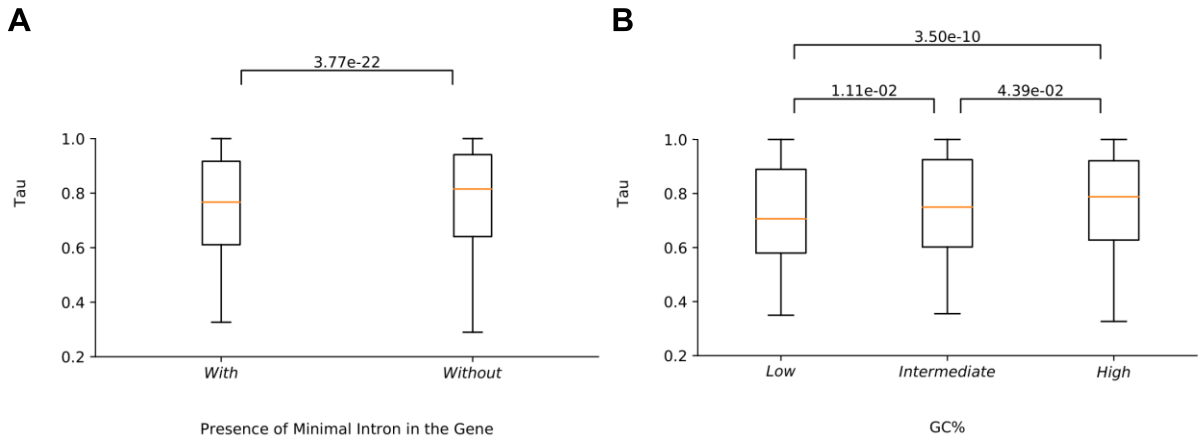


Figure 10.1 - Distribution of gene tissue-specificity expression of different groups of human genes. (a) Minimal intron-containing genes (n=4,459) show lower tissue specificity than genes without minimal introns (n=13,617). (b) Low GC% minimal introns (n=1,080, GC%<=43.09-5) have the lowest values of tissue specificity among the three groups of minimal intron-containing genes. High GC% group (n=2,886, GC%>=43.09+5), on the other hand, shows the highest values. Intermediate GC% minimal intron-containing genes (n=493, GC%=43.09±5) show values between the other two groups. Distributions of values in each population of genes are represented as boxplots. The coloured central line represents the median, the first quartile (25th percentile) by the bottom line of the box, the third quartile (75th percentile) by the upper line of the box, and the maximum and minimum values by the lines at the end of the whiskers. Outliers, if present, are represented by points. P-values from the non-parametric Mann–Whitney U test are shown above the horizontal bracket.

Source: By the author.

Even though the tissue specificity might vary between different groups of minimal intron-containing genes, there is a considerable proportion of genes with high tissue specificity in all groups. However, are the tissues responsible for high values of τ the same to low GC% and high GC% minimal intron-containing genes? To answer this question, we separated the population of genes with high tissue specificity (τ>=0.7) in each group of minimal intron-containing genes. We then used the tissue of higher expression to classify each gene. We verified that a large portion of genes tends to be classified in a few tissues. For example, 25.36% of genes with minimal introns of low GC% and high τ have the testis as the tissue with higher expression. Interestingly, we also verified that the proportion of genes with minimal introns with high GC% and high τ that have testis as the most expressed tissue is significantly lower than that of low GC% minimal intron-containing genes (20.97% x 25.36%, respectively, p-value 3.30e-02).

To allow a more comprehensive analysis, we compared this proportion for the two classes of genes in all tissues by subtracting the results of low GC% and high GC% minimal intron-containing genes. Values were subtracted in order to avoid overrepresentation of specialized tissues. In this way, positive values indicate that a certain tissue is more enriched with low GC% minimal intron-containing genes than with high GC% minimal intron-containing genes. We found a significant difference in this representation in several other tissues (Table 10.2). Interestingly, cultured cells and testis were the samples with the highest difference between proportions for low and high GC% minimal intron-containing genes (Table 10.2). We consider that this result is consistent with our finding that genes with low GC% minimal intron are enriched in functions related to cell division since cultured cells present high mitotic rates and testis display a high meiotic rate. High GC% minimal intron-containing genes, on the other hand, are enriched by tissues with low division and high metabolic rates (Table 10.2). In fact, cellular division is absent in red blood cells, which are the most abundant cell type in the Whole Blood tissue, [146] and spleen is a highly metabolic organ responsible for the phagocytosis of erythrocytes, recycling of iron and destruction of blood-borne pathogens. [147]

Table 10.2 - Fraction of genes with high tissue-specific expression (**τ**>=0.7) from low GC% (L_i) and high GC% (H_i) minimal intron-containing genes populations according to the tissue of higher expression.

Tissue/Cell [a]	L_i (%)	H_i (%)	L_i - H_i (%)	P-value
EBV-transformed lymphocytes	16.19	6.70	9.48	1.02e-11
Testis	25.36	20.97	4.39	3.30e-02
Cultured fibroblasts	6.83	2.65	4.18	6.59e-06
Brain - Cerebellum	1.80	7.94	-6.15	4.50e-07
Whole Blood	1.26	5.08	-3.82	1.33e-04
Spleen	1.44	3.94	-2.51	6.15e-03

[a] Tissues/cells corresponding to the three highest/lowest values of L_i - H_i are shown.
Source: By the author.

10.3 Population of low GC% minimal intron-containing genes is enriched with oncogenes

We verified that the population of genes containing minimal intron of low GC% are enriched in genes related to cell division and highly expressed in cultured cells that display high mitotic rates. Cancer cells are also known to display high mitotic rates and therefore it is reasonable to suppose that this set of genes may also play a considerable role in this context. To verify this hypothesis, we used a catalogue of genes which somatic mutations have been previously associated with a documented activity that promotes oncogenic transformation. [148] Using this list of oncogenic genes, the proportion of genes with and without minimal introns that have been classified as oncogenes were calculated. It was observed that minimal intron-containing genes have a greater fraction of oncogenes within its population than genes without minimal introns (Figure 10.2a). Later, the proportion of low, intermediate and high GC% minimal intron-containing genes classified as oncogenes were computed. Interestingly, the population of low GC% minimal intron-containing genes has the greatest fraction of genes

classified as oncogenes than in other minimal intron-containing genes populations (Figure 10.2b). To verify if the observed proportion of oncogenes within genes populations was significantly different from each other, the observed number of genes of each group in the oncogenic list was compared to the number of genes of those groups not classified as oncogenes using the Chi-square test of independence (Figure 10.2).

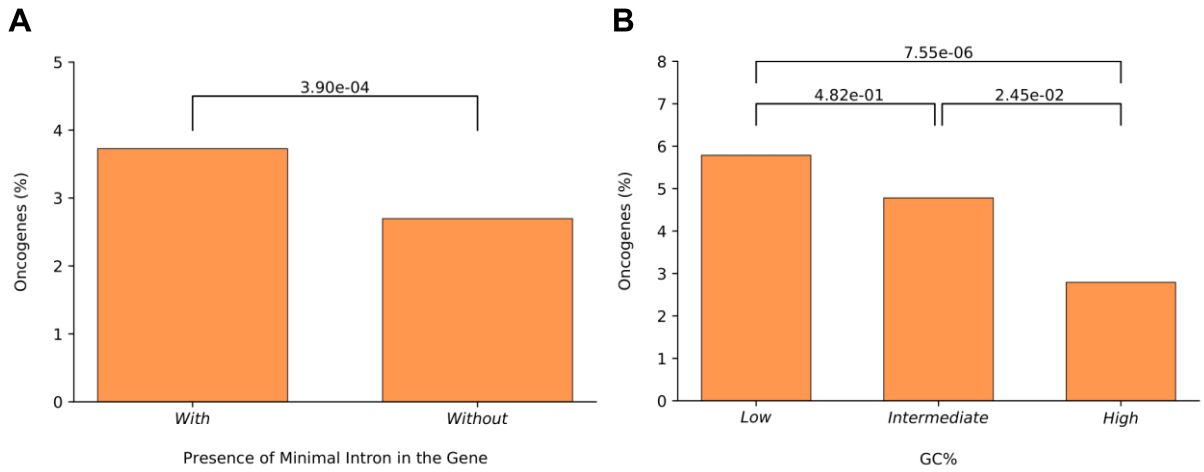


Figure 10.2 - Proportion of oncogenes within different groups of human genes. (a) Genes with minimal introns show a higher proportion of genes which mutations have been correlated with oncogenic transformation than genes without minimal introns. (b) Among minimal intron-containing genes, the low GC% group shows the highest proportion of cancer-related genes among its population. Minimal intron-containing genes were divided into three populations based on the GC content of the minimal introns: low GC% (<=43.09-5), intermediate GC% (43.09±5), and high GC% (>=43.09+5). P-values from the Chi-square test of independence calculated comparing the number of genes of each group related to cancer and not related to cancer are shown above horizontal brackets.

Source: By the author.

Is noteworthy that cancer is fundamentally related to lack of control in cell division and that it has been previously described that intron retention is the form of splicing that most differs between normal and cancerous tissues. [75] Curiously, our previous analyses have pointed to the functional importance of low GC% minimal intron-containing genes to cell division. Our analysis has also related the increase of GC content of minimal introns to an increase in the percentage of intron retention events. Therefore, it is possible to hypothesize that mutations within low GC% minimal intron-containing genes could result in the increase of the GC content of these genes that might lead to more intron retention events that in turn would result in diminished

control of the division cellular process. In that scenario, a strong negative selection against mutations that increase the GC content of low GC% minimal introns derived from genes related to cell division would occur to avoid excessive cell proliferation. Therefore, this hypothesis would explain the enrichment of oncogenes within the population of low GC% minimal intron-containing genes.

10.4 Low GC% minimal intron-containing genes are highly expressed in the meiotic stages of spermatogenesis

Our analysis of enrichment of gene ontology terms from low GC% minimal intron-containing genes also showed enrichment of the “Meiotic cell cycle” term (Table 10.1). Moreover, we observed that a large fraction of those genes showed high tissue-specific expression in the human testis (Table 10.2). In this section, the relation of low GC% minimal intron-containing genes and meiotic division will be better investigated by analysing the expression profile of these genes in human germ cells derived from different steps of the spermatogenesis process.

The spermatogenesis is an intricate cell differentiation process in which haploid sperm cells are formed from diploid spermatogonial stem cells. This complex process is composed of a pre-meiotic phase, followed by a proper meiotic phase and later by the post-meiotic phase. In this analysis, the expression profile of different types of genes was retrieved from RNA-Seq data from a previous study involving six cells of all three phases: Adark and Apale spermatogonia from the pre-meiotic phase; spermatocytes during leptotene/zygotene, early and late pachytene from prophase I of meiotic phase; and round spermatid from the post-meiotic phase (Figure 10.3a). [149] Interestingly, it has been reported that transcription is not a continuously active process during all stages of spermatogenesis. There is a significant reduction in the number of transcripts from Apale spermatogonia to leptotene/zygotene spermatocyte, followed by a transcription increase during early pachytene spermatocyte and a subsequent decrease upon later stages of spermatogenesis. [149]

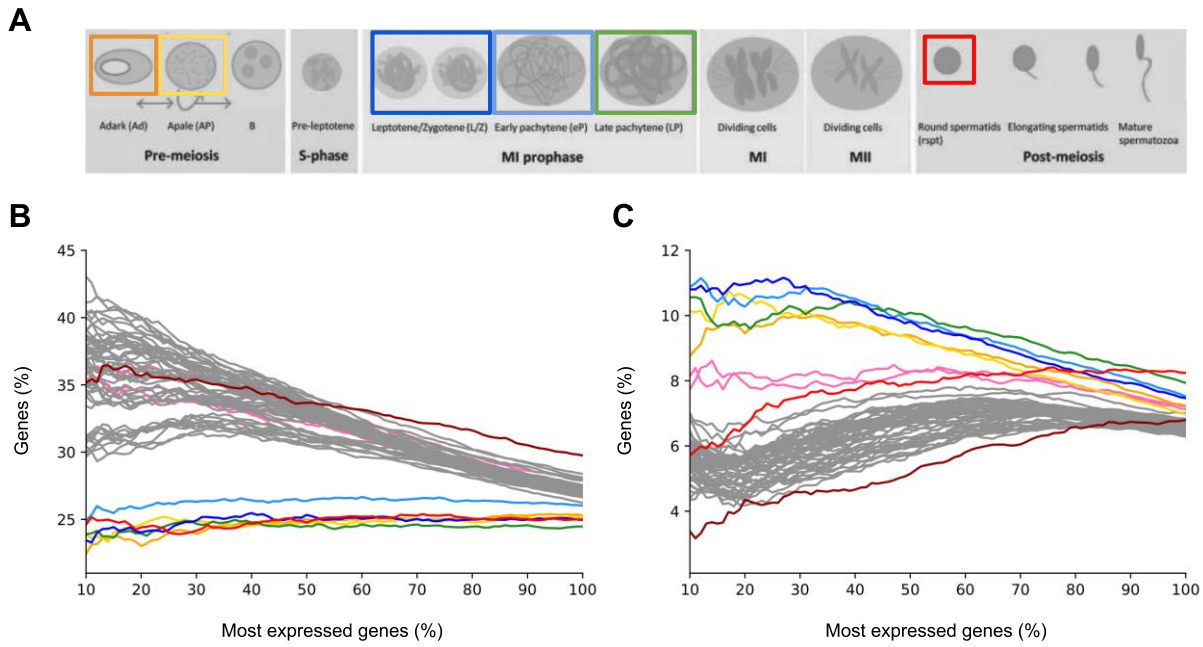


Figure 10.3 - The proportion of human genes within the most expressed genes in different human tissues/cells. For this analysis, different human germ cells of different stages of spermatogenesis were analysed: Adark and Apale spermatogonia from the pre-meiotic phase; spermatocytes during leptotene/zygotene, early and late pachytene from prophase I of meiotic phase; and round spermatid from the post-meiotic phase (a). The proportion of minimal intron-containing genes and low GC% minimal intron-containing genes among the most expressed genes are shown in (b) and (c) respectively. X-axis shows the percentage of most expressed genes considered for calculating the proportion of presence of minimal intron or low GC% minimal intron-containing genes. Cells of spermatogenesis show a lower proportion of minimal intron-containing genes among the most expressed genes in relation to other cells/tissues (b). Germ cells of the meiotic phase show a high proportion of low GC% minimal intron-containing genes in the most expressed genes, followed by cells from the pre-meiotic phase of spermatogenesis, and mitotically dividing cells. Post-meiotic germ cell, with no division, shows a similar proportion to other tissues. Whole blood, a tissue composed mainly by cells incapable of division, shows the lowest proportion of all tissues/cells (c). Line colours of germ cells correspond to the box colours in (a), mitotically dividing cultured fibroblasts and EBV-transformed lymphocytes are shown in pink, whole blood in dark red, other tissues in grey. Genes with minimal introns with GC%<= 43.09-5 were classified as low GC% minimal intron-containing genes. Only genes with Transcripts Per Million (TPM) greater than 1 were analysed.
Source: (a) Adapted from JAN *et al*.; [149] (b, c) By the author.

In order to understand the dynamic expression of minimal intron-containing genes during spermatogenesis, all human protein-coding genes were ranked according to their expression in each of the six analysed germ cells. Then, we calculated the proportion of those genes among diverse percentiles of expression. In order to compare our results, the same was done to gene expression data from not-dividing and mitotically dividing human tissues/cells retrieved from the Genotype-Tissue Expression (GTEx) database. Interestingly, the proportion of minimal intron-

containing genes among the most expressed genes is remarkably lower in the germ cells when compared to what is observed in the other human tissues/cells (Figure 10.3b).

This observation of a lower proportion of minimal-intron-containing genes among highly expressed genes from germ cells could be explained by a series of reported events that have been related to the IR phenomena. Firstly, we should remember that IR is commonly associated with introns of a short size where splicing occurs via intron definition. [66] Experiments involving heat shock have demonstrated that intron-retained transcripts could be kept in the nucleus for future export after further processing. [71] This strategy prevents the deleterious effects of erroneous transcripts that could lead to the production of aberrant proteins if those transcripts are exported to the cytoplasm. Together, we could speculate that high expression of genes with minimal introns could result in a higher proportion of transcripts with retained introns in the cell as minimal introns are very short. In cells that are not dividing, the cell could cope with this situation by adopting the same strategy that cells use in case of heat shock, keep intron-retained transcripts in the nucleus. However, if the cell is dividing, as is the case of germ cells, the nucleus no longer exists. In this way, the lower proportion of minimal intron-containing genes among the most expressed genes may be the strategy adopted by the germs cells in order to avoid an exacerbate number of transcripts with retained introns. Furthermore, the germs cells have an intricate system involving transcripts with retained introns that are stored in polyribosomes for posterior expression of those transcripts during the silent stages of spermatogenesis. [72] In this way, a disruption of this system by overloading it with transcripts with retained introns might have severe consequences. Therefore, it is possible to speculate the existence of a strong negative selection for highly expressed minimal intron-containing genes, in a way that only minimal intron-containing genes that are essential to the spermatogenesis process could have high levels of expression.

Interestingly, although minimal intron-containing genes are scarce in the population of highly expressed genes of human germ cells, if the same analysis is repeated considering only the proportion of low GC% minimal intron-containing genes, we observe that they are more representative in highly expressed genes of germ cells than in other human tissues/cells (Figure 10.3c). Considering that our previous gene

ontology enrichment analysis showed that this class of genes is enriched in genes involved in the meiotic cell cycle, chromosome segregation and regulation of cell cycle process, we understand that they are performing pivotal roles in meiosis. Moreover, as mentioned early, it has been reported that IR program is an important component of the meiotic cell, being responsible for the storage of transcripts in polyribosomes that ensures a longer half-life of these transcripts compared to transcripts with no retention. This mechanism compensates the lack of transcription during part of the meiosis, due to the chromatin condensation. [72] In this context, the IR observed in the highly expressed genes with low GC% minimal intron that have been related to meiosis might be a desirable feature to this mechanism of spermatogenesis and not a negatively selected feature as is observed for high GC% minimal intron containing-genes that have not been functional related to meiosis.

Moreover, if availability of transcripts from genes with minimal introns with low GC% is indeed regulated by IR, this would imply that tight regulation of such events would be an important factor in meiosis. In that scenario, a possible increase of GC content could be negative since it might induce an increase in retention levels. That would provide a possible explanation of why such genes remained at low GC% levels despite the increase of body temperature.

Interestingly, our analysis also shows that cells of the meiotic phase of spermatogenesis (leptotene/zygotene, early and late pachytene spermatocytes) have a greater proportion of low GC% minimal intron containing-genes among the highly expressed genes than pre-meiotic cells (Adark and Apale spermatogonia). It is important to notice that gene expression of our analysis is an indirect measure of the number of transcripts found in the analysed cell. Therefore, what is actually observed is a higher proportion of the transcripts derived from low GC% minimal intron-containing genes in cells of the meiotic phase of spermatogenesis (Figure 10.3c). These cells have condensed chromosomes and thus transcription must be somewhat impaired. In this way, we speculate that the higher proportion of transcripts from low GC% minimal intron-containing genes in these cells is due to the lower degradation rate of these transcripts as the actual transcription rate must already be depressed. This provides further support to our hypothesis of the importance of the IR within low GC% minimal intron-containing genes.

Is also interesting to notice that, despite the fact that Apale spermatogonia undergoes mitosis, cells from the pre-meiotic stage of spermatogenesis show a higher proportion of low GC% minimal intron-containing genes in relation to other mitotically dividing-cells (shown in pink in Figure 10.3c). This is another evidence that the observed transcripts from low GC% minimal intron-containing genes of meiotic germ cells were transcribed early during the pre-meiotic phase. It is worth mentioning that mitotically dividing-cells, on the other hand, show a higher proportion of low GC% minimal intron-containing genes when compared to not-dividing tissues/cells (Figure 10.3c). This corroborates to the functional importance of the low GC% minimal intron-containing genes to the division process as suggested by our previous analysis. This observation is also true to the round spermatid cell (red line), a germ cell from the post-meiosis stage that does not divide and is able to resume transcription. [150] In this cell, the genetic material has already been segregated and the nuclear membrane reconstituted. The round spermatid cell shows a similar proportion of low GC% minimal intron-containing genes to the one observed in other human tissues (grey lines, Figure 10.3c). The relation between low GC% minimal intron-containing genes is once more supported by the observation of its lower representation in the most expressed genes found in the whole blood (Figure 10.3c, dark red line). This tissue is mainly composed of red blood cells that are incapable of division. [146]

Taken together, these observations show that mitotic cells display an intermediate position between meiotic related cells and other tissues in terms of abundance of genes with minimal introns of low GC% (Figure 10.3c, pink lines). In addition, analysis of the fraction of low GC% minimal intron-containing genes that make up only the 10% most expressed genes also confirm that those differences in proportion between pre-meiotic/meiotic, mitotic and other tissues are statistically significant (Figure 10.4). Therefore, it is possible to hypothesize that, although mitotic cells must also have a similar relationship with intron retention of low GC% minimal introns containing genes, it must be less extensive than that observed in pre-meiotic/meiotic cells.

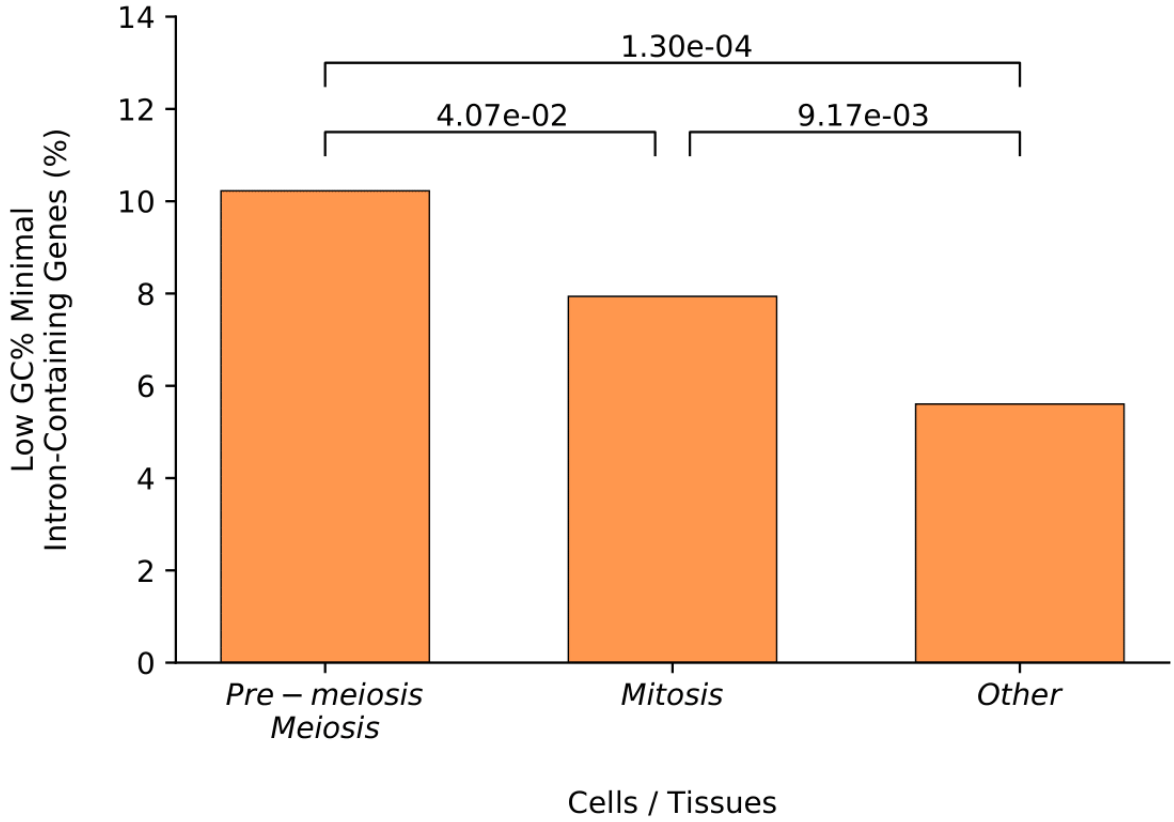


Figure 10.4 - The mean fraction of low GC% minimal intron-containing genes among the 10% most expressed genes of pre-meiotic/meiotic, mitotic and other tissues/cells. Cells involved in pre-meiosis/meiosis show the highest representation of low GC% minimal intron-containing genes, followed by cells involved in the mitotic division. P-values from the non-parametric Mann–Whitney U test are shown above the horizontal bracket.
Source: By the author.

10.5 Materials and methods

10.5.1 Gene ontology analysis of genes with low and high GC% minimal introns

Human genes with minimal introns were classified into two groups according to the GC content of its harboured minimal introns. It was defined as low GC% those minimal introns with GC%<=43.98, and as high GC% those with GC%>43.98. Genes with minimal introns classified in more than one group were removed from the analysis. The analysis that verifies any enrichment of gene ontology terms in each gene population was conducted using the Panther software (version 14.1). [145] A list of all human protein-coding genes was used as a reference list. The analysis was conducted using Fisher's Exact test followed by the Bonferroni correction of multiple comparisons. The human genome version GRCh37.p13 was used in this analysis.

10.5.2 Tissue specificity of different groups of genes

In this analysis, a dataset with gene median level of expression per tissue, measured by transcripts per million (TPM), was obtained from a processed RNA-Seq data of GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) of the Genotype-Tissue Expression (GTEx) Project. [151] From median TPM values of genes per tissue, values of Tau (τ) were calculated for each gene. Between many methods, Tau was chosen as it is considered the most robust metric for measuring tissue specificity. [152] Tau is defined as:

$$\tau = \frac{\sum_{i=1}^{n}(1-\widehat{x_i})}{n-1};\ \widehat{x_i} = \frac{x_i}{\max\limits_{1\le i\le n}(x_i)}.$$

Where n is the number of analysed tissues and x_i is the expression of the gene in the tissue i. Tau values range between 0 to 1, where 0 means that the same expression profile of a gene is observed in all analysed tissues, whereas 1 refers to a gene which is expressed in only one of the analysed tissues.

Using the GRCh38.p13 version (NCBI accession number GCA_000001405.28) of the human genome, [142] genes were classified based on the presence of minimal introns (introns of 78 bp to 98 bp), and also according to the GC content of its harboured minimal introns. It was defined as low, intermediate and high GC% minimal intron-containing genes if the minimal intron has GC%<=43.09-5, GC%=43.09±5 and GC%>=43.09+5, respectively. Genes with minimal introns classified in more than one group were removed from the analysis. Distributions of Tau values for each type of genes were shown using boxplot charts and statistical analysis was done using the non-parametric Mann–Whitney U test. This and the following analysis were conducted in Python programming language (version 3.7.3) with support of Biopython (version 1.72), Pandas (version 0.23.4), Matplotlib (version 2.2.3), Seaborn (version 0.9.0) and Scipy (version 1.3.1) libraries unless stated.

10.5.3 Percentage of tissues within low GC% and high GC% minimal intron-containing genes with high values of tissue specificity

In this analysis, only low GC% and high GC% minimal intron-containing genes from the previous analysis with Tau values greater or equal to 0.7 were used. The name of the tissue where the gene was most expressed was obtained. Then, the percentage of appearance of each tissue was calculated for each group of genes. The percentage value of a tissue from high GC% minimal intron-containing genes (H_i) was subtracted from the value of the same tissue from low GC% minimal intron-containing genes (L_i) to avoid overrepresentation of specialized tissues. Positive values of L_i - H_i indicate that a certain tissue is more enriched with low GC% minimal intron-containing genes than with high GC% minimal intron-containing genes. Statistical analysis was done using the Chi-square test of independence. For this analysis, Matplotlib (version 2.2.3) and Seaborn (version 0.9.0) libraries were not used.

10.5.4 Percentage of low GC% minimal intron-containing genes related to cancer

For this analysis, a catalogue of genes which mutations have been causally implicated in cancer was retrieved from The Cancer Gene Census (CGC) portal. [148] We classified as oncogenes only those genes with documented activity relevant to cancer, with reported mutation that causes gene production changes that promotes oncogenic transformation (genes classified as “Tier 1” in the CGC dataset). We also filtered genes based on their role in cancer, only genes described as “oncogene” were used. In this analysis, we used the GRCh38.p13 version of the human genome. [142] Firstly, the proportion of genes with or without minimal intron (introns of 78 bp to 98 bp) were analysed. Later, the proportion of low, intermediate, and high GC% minimal intron-containing genes were calculated. The observed number of genes of each group in the filtered oncogenic list was then compared to the number of those groups within genes not related to cancer. It was defined as low, intermediate and high GC% minimal intron-containing genes if the minimal intron has GC%<=43.09-5, GC%=43.09±5 and GC%>=43.09+5, respectively. Genes with minimal introns classified in more than one group were removed from the analysis. Statistical analysis was done using the Chi-square test of independence. This analysis was also supported by Numpy (version 1.16.4) library.

10.5.5 Proportion of minimal intron-containing genes within the most expressed genes in different human tissues/cells

In this analysis, we used the GRCh38.p13 version of the human genome. [142] Data from RNA-Seq experiments of Adark spermatogonia, Apale spermatogonia, leptotene/zygotene spermatocyte, early spermatocyte, late pachytene spermatocyte and round spermatid were retrieved from NCBI (SRA) under the accession number SRP069329. Raw RNA-Seq reads were downloaded using SRA toolkit (version 2.9.6), trimmed using Trimmomatic software (version 0.39) and mapped to the human genome with Hisat2 software (version 2.1.0). Using the Sequence Alignment/Map (SAM) output file, the number of reads mapped to each gene was counted using HTSeq (version 0.11.2). The expression level of each gene, measured in normalized counts per million (CPM), was then calculated using R programming language (version 3.6.1) supported by limma (version 3.42.0) and edgeR (version 3.28.0) packages. CPM values, together with the length of the gene defined as the length of overlapped exons, were transformed to TPM metric.

Genes were divided into those without and with minimal introns (introns of 78 bp to 98 bp) and further classified as low, intermediate and high GC% minimal intron-containing genes if the minimal intron has GC%<=43.09-5, GC%=43.09±5 and GC%>=43.09+5, respectively. Genes with minimal introns classified in more than one group were removed from the analysis. Only genes with TPM values greater than 1 were used in this analysis to filter out low signal-to-noise data. The proportion of each group of genes between different percentiles of the most expressed genes in the germ cells was calculated. For comparison purposes, the proportion of each group of genes was also calculated from median TPM data of other human tissues, previously obtained from a processed RNA-Seq data of GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) of the Genotype-Tissue Expression (GTEx) Project. [151] Later, we also calculated the proportion of low GC% minimal intron-containing genes from the universe of the 10% most expressed genes from tissues classified as “pre-meiotic/meiotic”, “mitotic” and “other”. Statistical analysis was done using the non-parametric Mann–Whitney U test. This analysis was also supported by Numpy (version 1.16.4) library.

Chapter 11

CONCLUSION

Analyses of genomes from Vertebrata phylum show unique features of minimal introns from this group. Firstly, comparison of introns sizes in deuterostomes indicates that vertebrate introns display a narrow peak close to the minimum size that is not seen in other phyla of deuterostomes, such as Echinodermata, Hemichordata and Cephalochordata. We speculate that this peak could be explained by the presence of a splicing machinery in the Vertebrata that is responsible for the evolutionary pressure to keep introns at a short size that is absent in many other deuterostomes. Furthermore, even though minimal introns from vertebrates tend to be larger than that observed in other eukaryotes, the modal length is conserved from lower to higher vertebrates (Figure 8.4). This suggests a strong selective force acting in the maintenance of an optimal intron size that may be related to a conserved splicing machinery within vertebrates.

Our studies with minimal introns of vertebrates suggest that increase in the body temperature has a great influence on minimal introns GC content. The analyses using the human genome have shown that minimal intron-containing genes tend to be responsive to increases in temperature, assuming a bimodal distribution of low and high GC contents. Such a trend is not observed in genes that do not harbour minimal introns, but instead, a unimodal distribution with a peak at low GC% is observed (Figure 9.7). We propose that genes with minimal introns may have a different splicing machinery from those genes that do not harbour minimal intron.

How the mRNA processing of minimal intron-containing genes is affected by temperature is still a question to solve. However, what we know is that part of these genes may have coped with the increase of the body temperature by increasing its GC content. The other part of the population of these genes could not adopt this solution

as the transition from low to high GC% might be a cumbersome process due to increase of IR events. This increase of IR levels may be too deleterious for genes related to basal cellular functions such as cell division, mRNA splicing and cell cycle checkpoint. In fact, we observed that mutations in low GC% minimal introns are linked to oncogenic transformation. We intend, in future analysis, to better explore the relationship between cancer and these low GC% minimal intron-containing genes. Using a database of described mutations correlated to cancer, retrieved from genome-wide association studies (GWAS), we intend to analyse if these mutations tend to increase the GC content of low GC% minimal introns that could explain the increase of intron retention observed in cancerous tissues.

Another interesting question is: how do low GC% minimal introns cope with an increase in body temperature? We speculate that the negative consequences of temperature increase in low GC% minimal introns are not too harmful than the consequences of changing its sequence towards high GC% in somatic cells. However, when it comes to cells involved with the perpetuation of the species this negative pressure of increased temperature becomes more critical. In this way, a solution should be adopted. Funny as it may seem, we proposed that mammalians have decided to reduce the temperature of the cells where meiosis occurs as a solution. Males have testes outside the body cavity where the temperature could be kept at 2 ^{o}C to 8 ^{o}C lower than the rest of the body in mice, humans and bulls. [153] In humans, a study with 150 infertile individuals shows a mean scrotal temperature of 0.4 °C for the right and 0.5 °C for the left higher than fertile individuals. [154] This indicates how harmful the temperature increase is to the meiotic system. In fact, the germ cells of meiosis I spermatocytes are one of the most vulnerable cells to heat stress in humans and rats. [155-156] The spermatogenesis suppression due to increased temperature is reported to be caused by a greater number of apoptosis events in germ cells that in turn is caused by increased DNA fragmentation related to defected repair mechanisms. [157] As low GC% minimal intron does not seem to have a specific solution to temperature increase, we propose that, when testis has elevated temperature, these introns may experience processing errors. Interestingly, low GC% minimal intron-containing genes are related to DNA repair (data not shown) and thus any defect on these genes could explain the increase in apoptosis due to high levels of DNA fragmentation. Furthermore, increased testis temperature has also been correlated to mRNA degradation [157]. Once more,

genes with low GC% minimal introns is related to functions involving mRNA stabilization.

Females mammals have also seemed to adopt the solution of decreasing the temperature of crucial organs where meiosis occurs as most of the males did. In species such as cattle, pigs, rabbits and humans, ovaries are kept 1 °C to 1.5 °C lower than the rest of the body. Differently from the meiotic process observed in males, the females have a limited number of meiotic cells and the meiosis is not a continuous process. The female oogenesis shows meiotic arrest during the primary oocyte that is arrested in metaphase I and secondary oocyte in metaphase II. The transition to metaphase I to metaphase II is stimulated by the luteinizing hormone (LH) during ovulation and this process is known as oocyte maturation. Later, the secondary oocyte exits meiotic arrest with the transient increase of intracellular calcium ion at the time of fertilization. [158-159] The temperature increase has been reported to be deleterious to oocyte maturation and to impede the division of secondary oocyte even after fertilization. [153,160] Interestingly, low GC% minimal intron-containing genes have been related to cell cycle checkpoint, a regulatory mechanism that blocks cell cycle progression. We once again propose that errors in oogenesis may be related to defects on low GC% minimal intron-containing genes due to temperature increase.

One might wonder, however, how have birds coped with high body temperature effects in spermatogenesis as the testis are kept inside the abdominal cavity? In this scenario, it is noteworthy that the influence of temperature in spermatogenesis is not the only strong evolutive pressure that birds are subjected to, aerodynamic streamlining is crucial on bird evolution. The thermotolerance observed by bird's spermatogenesis is reported to be related to increased polyadenylation of Hsp70 and ubiquitin transcripts. [161] Furthermore, differences in the heat shock protein A2 (HSPA2) from birds and mammals may be caused by differential selection pressures in these two groups. Adaptive changes in birds HSPA2 are likely to be temperature-driven and result in a protein that could adapt to elevated temperature with an advantage in response to heat stress. The HSPA2 in mammals, however, is unlikely to function efficiently during prolonged exposure to high temperature. [162] These observations corroborate our hypothesis of disruption in the mRNA processing of low GC% minimal intron-containing genes during the elevated temperature of spermatogenesis. As a

result, more defected proteins may be produced that leads to ubiquitination or even to the chaperone pathway due to misfolding.

Curiously, some mammals such as the cetaceans and pinnipeds resemble birds by having the testis inside the body cavity considered as a morphological adaptation of streamlining and axial swimming style. The locomotion style of these species may also impact the thermoregulation of the female reproductive system that is surrounded by thermogenic muscle and insulating blubber. [163] Similarly to other mammals, cetaceans and pinnipeds do not cope with an increase in temperature of reproductive tissues. In those animals, for both females and males, a countercurrent system was adopted as a solution to cool reproductive organs by carrying cool blood from the surface of the body, such as fins. [164–166] In this way, spermatogenesis and oogenesis may occur in lower temperature than the rest of the body.

All these observations point to the presence of a thermosensitive splicing mechanism responsible for the processing of introns from minimal intron-containing genes. Our study opens new questions on how the temperature of high body temperature vertebrates may influence the evolution dynamic of genomes. We propose a methodology for detecting the thermosensitive genes by the use of minimal introns as proxies. Furthermore, although our results have shown that some genes have coped with elevated temperature by increasing the GC content of its sequences, it is still unknown the reason for the GC variation and why low GC sequences show detrimental effects in high temperature systems. The mechanism behind the proposed hypothesis is still not solved. In addition, the relation of low GC minimal intron-containing genes with cell division could represent a new interesting system for studying diseases related to division defects, such as cancer and infertility. As IR has also been observed to be an important factor to minimal introns GC variation, the participation of minimal intron-containing genes in diseases in which increase of IR could be associated, such as diabetes type I and cancer, [75,167] could be explored by this new perspective.

REFERENCES

1 KOONIN, E. V. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? **Biology Direct**, v. 1, p. 22, 2006. doi: 10.1186/1745-6150-1-22.

2 BERGET, S. M.; MOORE, C.; SHARP, P. A. Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 8, p. 3171–3175, 1977. doi: 10.1073/pnas.74.8.3171

3 CHOW, L. T. *et al*. An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. **Cell**, v. 12, n. 1, p. 1–8, 1977. doi: 10.1016/0092-8674(77)90180-5.

4 GILBERT, W. Why genes in pieces? **Nature**, v. 271, n. 5645, p. 501, 1978. doi: 10.1038/271501a0.

5 FORD DOOLITTLE, W. Genes in pieces: were they ever together? **Nature**, v. 272, n. 5654, p. 581–582, 1978. doi: 10.1038/272581a0.

6 ROGOZIN, I. B. *et al*. Origin and evolution of spliceosomal introns. **Biology Direct**, v. 7, p. 11, 2012. doi: 10.1186/1745-6150-7-11.

7 THE NOBEL PRIZE IN PHYSIOLOGY OR MEDICINE 1993. Available from: https://www.nobelprize.org/prizes/medicine/1993/press-release. Accessible at: Apr. 26, 2020.

8 BERK, A. J. Discovery of RNA splicing and genes in pieces. **Proceedings of the National Academy of Sciences of the United States of America**, v. 113, n. 4, p. 801–805, 2016. doi: 10.1073/pnas.1525084113.

9 FERAT, J. L.; MICHEL, F. Group II self-splicing introns in bacteria. **Nature**, v. 364, n. 6435, p. 358-61, 1993. doi: 10.1038/364358a0.

10 KAINE, B. P.; GUPTA, R.; WOESE, C. R. Putative introns in tRNA genes of prokaryotes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 80, n. 11, p. 3309–3312, 1983. doi: 10.1073/pnas.80.11.3309.

11 REINHOLD-HUREK, B.; SHUB, D. A. Self-splicing introns in tRNA genes of widely divergent bacteria. **Nature**, v. 357, n. 6374, p. 173–176, 1992. doi: 10.1038/357173a0.

12 RAGHAVAN, R.; MINNICK, M. F. Group I introns and inteins: disparate origins but convergent parasitic strategies. **Journal of Bacteriology**, v. 191, n. 20, p. 6193-202, 2009. doi: 10.1128/jb.00675-09.

13 CECH, T. R. Self-splicing of group I introns. **Annual Review of Biochemistry**, v. 59, n. 1, p. 543-68, 1990. doi: 10.1146/annurev.bi.59.070190.002551.

14 MCNEIL, B. A.; SEMPER, C.; ZIMMERLY, S. Group II introns: versatile ribozymes and retroelements. **Wiley Interdisciplinary Reviews:** RNA, v. 7, n. 3, p. 341–355, 2016. doi: 10.1002/wrna.1339.

15 PYLE, A. M. Group II intron self-splicing. **Annual Review of Biophysics**, v. 45, p. 183- 205, 2016. doi: 10.1146/annurev-biophys-062215-011149.

16 LAMBOWITZ, A. M.; ZIMMERLY, S. Group II introns: mobile ribozymes that invade DNA. **Cold Spring Harbor Perspectives in Biology**, v. 3, n. 8, p. a003616, 2011. doi: 10.1101/cshperspect.a003616.

17 YOSHIHISA, T. Handling tRNA introns, archaeal way and eukaryotic way. **Frontiers in Genetics**, v. 5, p. 213, 2014. doi: 10.3389/fgene.2014.00213.

18 SCHMIDT, C. A.; MATERA, A. G. tRNA introns: presence, processing, and purpose. **Wiley Interdisciplinary Reviews:** RNA, v. 11, n. 3, p. e1583, 2019. doi: 10.1002/wrna.1583.

19 CHOREV, M.; CARMEL, L. The function of introns. **Frontiers in Genetics**, v. 3, p. 55, 2012. doi: 10.3389/fgene.2012.00055.

20 ROY, S. W.; GILBERT, W. The evolution of spliceosomal introns: patterns, puzzles and progress. **Nature Reviews:** genetics, v. 7, n. 3, p. 211–221, 2006. doi: 10.1038/nrg1807.

21 BLACK, D. L. Mechanisms of alternative pre-messenger RNA splicing. **Annual Review of Biochemistry**, v. 72, n. 1, p. 291-336, 2003. doi: 10.1146/annurev.biochem.72.121801.161720.

22 WATSON, J. D. *et al*. **Molecular biology of the gene**. 5th ed. San Francisco: Pearson Education, 2004.

23 VICENS, Q.; CECH, T. R. A natural ribozyme with 3′,5′ RNA ligase activity. **Nature Chemical Biology**, v. 5, n. 2, p. 97, 2009. doi: 10.1038/nchembio.136.

24 PALMER, J. D.; LOGSDON, J. M. The recent origins of introns. **Current Biology**, v. 1, n. 4, p. 470-7, 1992. doi: 10.1016/0960-9822(92)90426-b.

25 STEGEMANN, S. *et al*. High-frequency gene transfer from the chloroplast genome to the nucleus. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 15, p. 8828–8833, 2003. doi: 10.1073/pnas.1430924100.

26 BRANDVAIN, Y.; WADE, M. J. The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. **Genetics**, v. 182, n. 4, p. 1129–1139, 2009. doi: 10.1534/genetics.108.100024.

27 LYNCH, M. The evolution of spliceosomal introns. **Current Opinion in Genetics & Development**, v. 12, n. 6, p. 701-10, 2002. doi: 10.1016/s0959-437x(02)00360-x.

28 DE CONTI, L.; BARALLE, M.; BURATTI, E. Exon and intron definition in pre-mRNA splicing. **Wiley Interdisciplinary Reviews:** RNA, v. 4, n. 1, p. 49-60, 2013. doi: 10.1002/wrna.1140.

29 PATEL, A. A.; STEITZ, J. A. Splicing double: insights from the second spliceosome. **Nature Reviews:** molecular cell biology, v. 4, n. 12, p. 960–970, 2003. doi: 10.1038/nrm1259.

30 WILL, C. L.; LUHRMANN, R. Spliceosome structure and function. **Cold Spring Harbor Perspectives in Biology**, v. 3, n. 7, p. a003707, 2011. doi: 10.1101/cshperspect.a003707.

31 SAINI, H. **Intron and small RNA localization in mammalian neurons**, 2019. (Doctoral Dissertation) - University of Massachusetts Medical School, 2019. doi: 10.13028/srmk-pk14.

32 TURUNEN, J. J. *et al*. The significant other: splicing by the minor spliceosome. **Wiley Interdisciplinary Reviews:** RNA, v. 4, n. 1, p. 61–76, 2013. doi: 10.1002/wrna.1141.

33 DIETRICH, R. C.; INCORVAIA, R.; PADGETT, R. A. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. **Molecular Cell**, v. 1, n. 1, p. 151-60, 1997. doi: 10.1016/s1097-2765(00)80016-7.

34 YU, J. *et al*. Minimal introns are not 'junk'. **Genome Research**, v. 12, n. 8, p. 1185-9, 2002. doi: 10.1101/gr.224602.

35 SAKHARKAR, M. K.; CHOW, V. T. K.; KANGUEANE, P. Distributions of exons and introns in the human genome. **In Silico Biology**, v. 4, n. 4, p. 387–393, 2004.

36 INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860-921, 2001. doi: 10.1038/35057062.

37 SINGH, J.; PADGETT, R. A. Rates of in situ transcription and splicing in large human genes. **Nature Structural & Molecular Biology**, v. 16, n. 11, p. 1128–1133, 2009. doi: 10.1038/nsmb.1666.

38 LYNCH, M. **The origins of genome architecture**. Sunderland: Sinauer Associates, 2007.

39 JO, B.-S.; CHOI, S. S. Introns: the functional benefits of introns in genomes. **Genomics & Informatics**, v. 13, n. 4, p. 112–118, 2015. doi: 10.5808/GI.2015.13.4.112.

40 SPIES, N. *et al*. Biased chromatin signatures around polyadenylation sites and exons. **Molecular Cell**, v. 36, n. 2, p. 245–254, 2009. doi: 10.1016/j.molcel.2009.10.008.

41 SCHWARTZ, S.; MESHORER, E.; AST, G. Chromatin organization marks exon-

intron structure. **Nature Structural & Molecular Biology**, v. 16, n. 9, p. 990–995, 2009. doi: 10.1038/nsmb.1659.

42 LAUDERDALE, J. D.; STEIN, A. Introns of the chicken ovalbumin gene promote nucleosome alignment in vitro. **Nucleic Acids Research**, v. 20, n. 24, p. 6589–6596, 1992. doi: 10.1093/nar/20.24.6589.

43 LIU, K. *et al*. Rat growth hormone gene introns stimulate nucleosome alignment in vitro and in transgenic mice. **Proceedings of the National Academy of Sciences of the United States of America**, v. 92, n. 17, p. 7724–7728, 1995. doi: 10.1073/pnas.92.17.7724.

44 JUNEAU, K. *et al*. Introns regulate RNA and protein abundance in yeast. **Genetics**, v. 174, n. 1, p. 511–518, 2006. doi: 10.1534/genetics.106.058560.

45 SHABALINA, S. A. *et al*. Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. **Molecular Biology and Evolution**, v. 27, n. 8, p. 1745–1749, 2010. doi: 10.1093/molbev/msq086.

46 BUCHMAN, A. R.; BERG, P. Comparison of intron-dependent and intron-independent gene expression. **Molecular and Cellular Biology**, v. 8, n. 10, p. 4395–4405, 1988. doi: 10.1128/mcb.8.10.4395.

47 OTTO, S. P.; BARTON, N. H. The evolution of recombination: removing the limits to natural selection. **Genetics**, v. 147, n. 2, p. 879–906, 1997.

48 SCHMUCKER, D. *et al*. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. **Cell**, v. 101, n. 6, p. 671–684, 2000. doi: 10.1016/s0092-8674(00)80878-8.

49 SCHWARTZ, S. H. *et al*. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. **Genome Research**, v. 18, n. 1, p. 88–103, 2008. doi: 10.1101/gr.6818908.

50 WANG, Z.; BURGE, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. **RNA**, v. 14, n. 5, p. 802-13, 2008. doi: 10.1261/rna.876308.

51 FURTH, P. A. *et al*. Sequences homologous to 5’ splice sites are required for the inhibitory activity of papillomavirus late 3' untranslated regions. **Molecular and Cellular Biology**, v. 14, n. 8, p. 5278–5289, 1994. doi: 10.1128/mcb.14.8.5278.

52 LEWIS, B. P.; GREEN, R. E.; BRENNER, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 1, p. 189–192, 2003. doi: 10.1073/pnas.0136770100.

53 GREEN, R. E. *et al*. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. **Bioinformatics**, v. 19, n. Suppl 1, p. i118–21, 2003. doi: 10.1093/bioinformatics/btg1015.

54 VALENCIA, P.; DIAS, A. P.; REED, R. Splicing promotes rapid and efficient mRNA export in mammalian cells. **Proceedings of the National Academy of Sciences of the United States of America**, v. 105, n. 9, p. 3386–3391, 2008. doi: 10.1073/pnas.0800250105.

55 SWINBURNE, I. A.; SILVER, P. A. Intron delays and transcriptional timing during development. **Developmental Cell**, v. 14, n. 3, p. 324–330, 2008. doi: 10.1016/j.devcel.2008.02.002.

56 TAKASHIMA, Y. *et al*. Intronic delay is essential for oscillatory expression in the segmentation clock. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 8, p. 3300–3305, 2011. doi: 10.1073/pnas.1014418108.

57 ALT, F. W. *et al*. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. **Cell**, v. 20, n. 2, p. 293–301, 1980. doi: 10.1016/0092-8674(80)90615-7.

58 EARLY, P. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. **Cell**, v. 20, n. 2, p. 313-9, 1980. doi: 10.1016/0092-8674(80)90617-0.

59 PAN, Q. *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. **Nature Genetics**, v. 40, n. 12, p. 1413–1415, 2008. doi: 10.1038/ng.259.

60 NILSEN, T. W.; GRAVELEY, B. R. Expansion of the eukaryotic proteome by alternative splicing. **Nature**, v. 463, n. 7280, p. 457–463, 2010. doi: 10.1038/nature08909.

61 TANG, S. J. *et al*. Cis- and trans-regulations of pre-mRNA splicing by RNA editing enzymes influence cancer development. **Nature Communications**, v. 11, n. 1, p. 799, 2020. doi: 10.1038/s41467-020-14621-5.

62 WARF, M. B.; BERGLUND, J. A. Role of RNA structure in regulating pre-mRNA splicing. **Trends in Biochemical Sciences**, v. 35, n. 3, p. 169–178, 2010. doi: 10.1016/j.tibs.2009.10.004.

63 GÓMEZ ACUÑA, L. I. *et al*. Connections between chromatin signatures and splicing. **Wiley Interdisciplinary Reviews:** RNA, v. 4, n. 1, p. 77–91, 2013. doi: 10.1002/wrna.1142.

64 MONTEUUIS, G. *et al*. The changing paradigm of intron retention: regulation, ramifications and recipes. **Nucleic Acids Research**, v. 47, n. 22, p. 11497–11513, 2019. doi: 10.1093/nar/gkz1068.

65 JACOB, A. G.; SMITH, C. W. J. Intron retention as a component of regulated gene expression programs. **Human Genetics**, v. 136, n. 9, p. 1043–1057, 2017. doi: 10.1007/s00439-017-1791-x.

66 LIM, L. P.; BURGE, C. B. A computational analysis of sequence features involved in recognition of short introns. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 20, p. 11193–11198, 2001. doi: 10.1073/pnas.201407298.

67 BERGET, S. M. Exon recognition in vertebrate splicing. **Journal of Biological Chemistry**, v. 270, n. 6, p. 2411-4, 1995. doi: 10.1074/jbc.270.6.2411.

68 MCGUIRE, A. M. *et al*. Cross-kingdom patterns of alternative splicing and splice recognition. **Genome Biology**, v. 9, n. 3, p. R50, 2008. doi: 10.1186/gb-2008-9-3-r50.

69 BRAUNSCHWEIG, U. *et al*. Widespread intron retention in mammals functionally tunes transcriptomes. **Genome Research**, v. 24, n. 11, p. 1774–1786, 2014. doi: 10.1101/gr.177790.114.

70 SAKABE, N. J.; DE SOUZA, S. J. Sequence features responsible for intron retention in human. **BMC Genomics**, v. 8, n. 1, p. 59, 2007. doi: 10.1186/1471-2164-8-59.

71 SHALGI, R. *et al*. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. **Cell Reports**, v. 7, n. 5, p. 1362–1370, 2014. doi: 10.1016/j.celrep.2014.04.044.

72 NARO, C. *et al*. An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. **Developmental Cell**, v. 41, n. 1, p. 82–93.e4, 2017. doi: 10.1016/j.devcel.2017.03.003.

73 DING, F. *et al*. Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. **BMC Genomics**, v. 15, n.1 ,p. 431, 2014. doi: 10.1186/1471-2164-15-431.

74 LING, Y. *et al*. Pre-mRNA splicing repression triggers abiotic stress signaling in plants. **The Plant Journal:** for cell and molecular biology, v. 89, n. 2, p. 291–309, 2017. doi: 10.1111/tpj.13383.

75 DVINGE, H.; BRADLEY, R. K. Widespread intron retention diversifies most cancer transcriptomes. **Genome Medicine**, v. 7, n. 1, p. 45, 2015. doi: 10.1186/s13073-015-0168-9.

76 ADUSUMALLI, S. *et al*. Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease. **Aging Cell**, v. 18, n. 3, p. e12928, 2019. doi: 10.1111/acel.12928.

77 JIAYAN, W. *et al*. Systematic analysis of intron size and abundance parameters in diverse lineages. **Science China:** life sciences, v. 56, n. 10, p. 968–974, 2013. doi: 10.1007/s11427-013-4540-y.

78 PIOVESAN, A. *et al*. Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI gene databank. **DNA Research**, v. 22, n. 6, p. 495–503, 2015. doi: 10.1093/dnares/dsv028.

79 REUGELS, A. M. *et al*. Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of Drosophila hydei. **Genetics**, v. 154, n. 2, p. 759–769, 2000.

80 WANG, D.; YU, J. Both size and GC-content of minimal introns are selected in human populations. **PloS One**, v. 6, n. 3, p. e17945, 2011. doi: 10.1371/journal.pone.0017945.

81 KIMURA, M. **The neutral theory of molecular evolution.** Cambridge: Cambridge University Press, 1983. doi: 10.1017/CBO9780511623486.

82 ZHU, J. *et al*. A novel role for minimal introns: routing mRNAs to the cytosol. **PloS One**, v. 5, n. 4, p. e10144, 2010. doi: 10.1371/journal.pone.0010144.

83 CREMER, T. *et al*. Chromosome territories - a functional nuclear landscape. **Current Opinion in Cell Biology**, v. 18, n. 3, p. 307–316, 2006. doi: 10.1016/j.ceb.2006.04.007.

84 CAIRA, J. N.; TIMOTHY, D.; LITTLEWOOD, J. Worms, Platyhelminthes. **Encyclopedia of Biodiversity**, v. 5, p. 863-899, 2013. doi: 10.1016/b0-12-226865-2/00287-x.

85 LITTLEWOOD, D. T. J.; WAESCHENBACH, A. Evolution: a turn up for the worms. **Current Biology**, v. 25, n. 11, p. R457–60, 2015. doi: 10.1016/j.cub.2015.04.012

86 LITTLEWOOD, D. T.; BRAY, R. A. **Interrelationships of the Platyhelminthes.** New York: CRC Press, 2000.

87 OLSON, P. D. *et al*. Phylogeny and classification of the Digenea (Platyhelminthes: Trematoda). **International Journal for Parasitology**, v. 33, n. 7, p. 733-55, 2003. doi: 10.1016/s0020-7519(03)00049-3.

88 CDC - SCHISTOSOMIASIS - BIOLOGY. Atlanta, 2019. Available from: https://www.cdc.gov/parasites/schistosomiasis/biology.html. Accessible at: Mar. 23, 2020.

89 SCHISTOSOMIASIS. Geneva, 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/schistosomiasis. Accessible at: Mar. 25, 2020.

90 NELWAN, M. L. Schistosomiasis: life cycle, diagnosis, and control. **Current Therapeutic Research, Clinical and Experimental**, v. 91, p. 5–9, 2019. doi: 10.1016/j.curtheres.2019.06.001.

91 VERJOVSKI-ALMEIDA, S. *et al*. Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. **Nature Genetics**, v. 35, n. 2, p. 148–157, 2003. doi: 10.1038/ng1237.

92 CARNEIRO, M.; FERRAND, N.; NACHMAN, M. W. Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the european rabbit (Oryctolagus cuniculus). **Genetics**, v. 181, n. 2, p.

593–606, 2009. doi: 10.1534/genetics.108.096826.

93 HONG, X.; SCOFIELD, D. G.; LYNCH, M. Intron size, abundance, and distribution within untranslated regions of genes. **Molecular Biology and Evolution**, v. 23, n. 12, p. 2392–2404, 2006. doi: 10.1093/molbev/msl111.

94 BERRIMAN, M. *et al*. The genome of the blood fluke Schistosoma mansoni. **Nature**, v. 460, n. 7253, p. 352–358, 2009. doi: 10.1038/nature08160.

95 PUCKER, B.; BROCKINGTON, S. F. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. **BMC Genomics**, v. 19, n. 1, p. 980, 2018. doi: 10.1186/s12864-018-5360-z.

96 SIBLEY, C. R.; BLAZQUEZ, L.; ULE, J. Lessons from non-canonical splicing. **Nature Reviews:** genetics, v. 17, n. 7, p. 407–421, 2016. doi: 10.1038/nrg.2016.46.

97 FOUSER, L. A.; FRIESEN, J. D. Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. **Cell**, v. 45, n. 1, p. 81–93, 1986. doi: 10.1016/0092-8674(86)90540-4.

98 SÉRAPHIN, B.; KRETZNER, L.; ROSBASH, M. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5′ cleavage site. **The EMBO Journal**, v. 7, n. 8, p. 2533-8, 1988. doi: 10.1002/j.1460-2075.1988.tb03101.x.

99 LESSER, C. F.; GUTHRIE, C. Mutations in U6 snRNA that alter splice site specificity: implications for the active site. **Science**, v. 262, n. 5142, p. 1982–1988, 1993. doi: 10.1126/science.8266093.

100 COOLIDGE, C. J.; SEELY, R. J.; PATTON, J. G. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. **Nucleic Acids Research**, v. 25, n. 4, p. 888–896, 1997. doi: 10.1093/nar/25.4.888.

101 SICKMIER, E. A. *et al*. Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. **Molecular Cell**, v. 23, n. 1, p. 49–59, 2006. doi: 10.1016/j.molcel.2006.05.025.

102 MERCER, T. R. *et al*. Genome-wide discovery of human splicing branchpoints. **Genome Research**, v. 25, n. 2, p. 290–303, 2015. doi: 10.1101/gr.182899.114.

103 SPINGOLA, M. *et al*. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. **RNA**, v. 5, n. 2, p. 221–234, 1999. doi: 10.1017/s1355838299981682.

104 CHEN, X. *et al*. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. **Bioinformatics**, v. 24, n. 9, p. 1121–1128, 2008. doi: 10.1093/bioinformatics/btn088.

105 BAILEY, T. L.; ELKAN, C. Fitting a mixture model by expectation maximization to

discover motifs in biopolymers. **Proceedings of the International Conference on Intelligent Systems for Molecular Biology**, v. 2, p. 28-36, 1994.

106 TAGGART, A. J. et al. Large-scale analysis of branchpoint usage across species and cell lines. **Genome Research**, v. 27, n. 4, p. 639–649, 2017. doi: 10.1101/gr.202820.115.

107 SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

108 SCHNEIDER, T. D.; STEPHENS, R. M. Sequence logos: a new way to display consensus sequences. **Nucleic Acids Research**, v. 18, n. 20, p. 6097–6100, 1990. doi: 10.1093/nar/18.20.6097.

109 SCHNEIDER, T. D. et al. Information content of binding sites on nucleotide sequences. **Journal of Molecular Biology**, v. 188, n. 3, p. 415–431, 1986. doi: 10.1016/0022-2836(86)90165-8.

110 LOBRY, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. **Molecular Biology and Evolution**, v. 13, n. 5, p. 660–665, 1996. doi: 10.1093/oxfordjournals.molbev.a025626.

111 CROOKS, G. E. WebLogo: a sequence logo generator. **Genome Research**, v. 14, n. 6, p. 130-148, 2004. doi: 10.1101/gr.849004.

112 HASHIM, F. A.; MABROUK, M. S.; AL-ATABANY, W. Review of different sequence motif finding algorithms. **Avicenna Journal of Medical Biotechnology**, v. 11, n. 2, p. 130-148, 2019.

113 CONTI, E.; IZAURRALDE, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. **Current Opinion in Cell Biology**, v. 17, n. 3, p. 316-325, 2005. doi: 10.1016/j.ceb.2005.04.005.

114 MAQUAT, L. E. Nonsense-mediated mRNA decay: a comparative analysis of different species. **Current Genomics**, v. 5, n. 3, p. 175-190, 2004. doi: 10.2174/1389202043349453.

115 WEN, J. *et al*. Splicing-dependent NMD requires Prp17 in Saccharomyces cerevisiae. **BioRxiv**, v. 1, p. 149245, 2017. doi: 10.1101/149245.

116 GATFIELD, D. Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. **The EMBO Journal**, v. 22, n. 15, p. 3960-3970, 2003. doi: 10.1093/emboj/cdg371.

117 TIAN, M. *et al*. Nonsense-mediated mRNA decay in Tetrahymena is EJC independent and requires a protozoa-specific nuclease. **Nucleic Acids Research**, v. 45, n. 11, p. 6848-6863, 2017. doi: 10.1093/nar/gkx256.

118 LONGMAN, D. *et al*. Mechanistic insights and identification of two novel factors in

the C. elegans NMD pathway. **Genes & Development**, v. 21, n. 9, p. 1075-1085, 2007. doi: 10.1101/gad.417707.

119 BEHRINGER, M. G.; HALL, D. W. Selection on position of nonsense codons in introns. **Genetics**, v. 204, n. 3, p. 1239-1248, 2016. doi: 10.1534/genetics.116.189894.

120 MORPHEUS. Available from: https://software.broadinstitute.org/morpheus. Accessible at: Apr. 25, 2020.

121 FIELDS, C. Information content of Caenorhabditis elegans splice site sequences varies with intron length. **Nucleic Acids Research**, v. 18, n. 6, p. 1509-1512, 1990. doi: 10.1093/nar/18.6.1509.

122 MAQUAT, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. **Nature Reviews:** molecular cell biology, v. 5, n. 2, p. 89-99, 2004. doi: 10.1038/nrm1310.

123 SATOH, N.; ROKHSAR, D.; NISHIKAWA, T. Chordate evolution and the three-phylum system. **Proceedings of the Royal Society B:** biological sciences, v. 281, n. 1794, p. 20141729, 2014. doi: 10.1098/rspb.2014.1729.

124 BLAIR HEDGES, S.; KUMAR, S. **The timetree of life**. New York: Oxford University Press, 2009.

125 SIMAKOV, O. *et al*. Insights into bilaterian evolution from three spiralian genomes. **Nature**, v. 493, n. 7433, p. 526-531, 2013. doi: 10.1038/nature11696.

126 OHNO, S. **Evolution by gene duplication**. New York: Springer, 2013.

127 HOLLAND, P. W. *et al*. Gene duplications and the origins of vertebrate development. **Development:** supplement, p. 125-133, 1994.

128 GELFMAN, S. *et al*. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. **Genome Research**, v. 22, n. 1, p. 35-59, 2012. doi: 10.1101/gr.119834.110.

129 GREER, A. E.; LAZELL, J. D.; WRIGHT, R. M. Anatomical evidence for a counter-current heat exchanger in the leatherback turtle (Dermochelys coriacea). **Nature**, v. 244, n. 5412, p. 181, 1973. doi: 10.1038/244181a0.

130 HUTCHISON, V. H.; DOWLING, H. G.; VINEGAR, A. Thermoregulation in a brooding female indian python, python Molurus bivittatus. **Science**, 1966. doi: 10.1126/science.151.3711.694.

131 CAREY, F. G.; TEAL, J. M. Heat conservation in tuna fish muscle. **Proceedings of the National Academy of Sciences of the United States of America**, v. 56, n. 5, p. 1464-1469, 1966. doi: 10.1073/pnas.56.5.1464.

132 RUBEN, J. The evolution of endothermy in mammals and birds: from physiology to fossils. **Annual Review of Physiology**, v. 57, n. 1, p. 69-95, 1995. doi:

10.1146/annurev.ph.57.030195.000441.

133 AMIT, M. *et al*. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. **Cell Reports**, v. 1, n. 5, p. 543-556, 2012. doi: 10.1016/j.celrep.2012.03.013.

134 BERNARDI, G. *et al*. The mosaic genome of warm-blooded vertebrates. **Science**, v. 228, n. 4702, p. 953-958, 1985. doi: 10.1126/science.4001930.

135 DEUTSCH, M.; LONG, M. Intron-exon structures of eukaryotic model organisms. **Nucleic Acids Research**, v. 27, n. 15, p. 3219-3228, 1999. doi: 10.1093/nar/27.15.3219.

136 BERNARDI, G. Isochores and the evolutionary genomics of vertebrates. **Gene**, v. 241, n. 1, p. 3-17, 2000. doi: 10.1016/s0378-1119(99)00485-0.

137 IVELL, R. Lifestyle impact and the biology of the human scrotum. **Reproductive Biology and Endocrinology**, v. 5, n. 1, p. 15, 2007. doi: 10.1186/1477-7827-5-15.

138 SCHMITZ, U. *et al*. Intron retention enhances gene regulatory complexity in vertebrates. **Genome Biology**, v. 18, n. 1, p. 216, 2017. doi: 10.1186/s13059-017-1339-3.

139 KUDLA, G. *et al*. High guanine and cytosine content increases mRNA levels in mammalian cells. **PLoS Biology**, v. 4, n. 6, p. e180, 2006. doi: 10.1371/journal.pbio.0040180.

140 KUEHNER, J. N.; PEARSON, E. L.; MOORE, C. Unravelling the means to an end: RNA polymerase II transcription termination. **Nature Reviews:** molecular cell biology, v. 12, n. 5, p. 283-94, 2011. doi: 10.1038/nrm3098.

141 THE HUMAN PROTEOME IN TESTIS - THE HUMAN PROTEIN ATLAS. Available from: https://www.proteinatlas.org/humanproteome/tissue/testis#stratum. Accessible at: Apr. 10, 2020.

142 THE GENOME REFERENCE CONSORTIUM. Available from: https://www.ncbi.nlm.nih.gov/grc. Accessible at: Apr. 10, 2020.

143 KINSELLA, R. J.; KÄHÄRI, A.; HAIDER, S.; et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. **Database:** the journal of biological databases and curation, v. 2011, p. bar030, 2011. doi: 10.1093/database/bar030.

144 GENE EXPRESSION OMNIBUS - GEO - NCBI. Available from: http://www.ncbi.nlm.nih.gov/geo/. Accessible at: Apr. 29, 2020.

145 MI, H. *et al*. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. **Nucleic Acids Research**, v. 47, n. D1, p. D419–D426, 2019. doi: 10.1093/nar/gky1038.

146 COMPOSITION OF THE BLOOD | SEER TRAINING. Available from:

https://training.seer.cancer.gov/leukemia/anatomy/composition.html. Accessible at: Apr. 14, 2020.

147 MEBIUS, R. E.; KRAAL, G. Structure and function of the spleen. **Nature Reviews:** immunology, v. 5, n. 8, p. 606-616, 2005. doi: 10.1038/nri1669.

148 TATE, J. G.; BAMFORD, S.; JUBB, H. C.; et al. COSMIC: the catalogue of somatic mutations in cancer. **Nucleic Acids Research**, v. 47, n. D1, p. D941-D947, 2019. doi: 10.1093/nar/gky1015.

149 JAN, S. Z.; VORMER, T. L.; JONGEJAN, A.; et al. Unraveling transcriptome dynamics in human spermatogenesis. **Development**, v. 144, n. 20, p. 3659-3673, 2017. doi: 10.1242/dev.152413.

150 VIBRANOVSKI, M. D.; CHALOPIN, D. S.; LOPES, H. F.; LONG, M.; KARR, T. L. Direct evidence for postmeiotic transcription during Drosophila melanogaster spermatogenesis. **Genetics**, v. 186, n. 1, p. 431-433, 2010. doi: 10.1534/genetics.110.118919.

151 GTEX PORTAL. Available from: https://gtexportal.org/home/datasets. Accessible at: Apr. 16, 2020.

152 KRYUCHKOVA-MOSTACCI, N.; ROBINSON-RECHAVI, M. A benchmark of gene expression tissue-specificity metrics. **Briefings in Bioinformatics**, v. 18, n. 2, p. 205-214, 2017. doi: 10.1093/bib/bbw008.

153 TAKAHASHI, M. Heat stress on reproductive function and fertility in mammals. **Reproductive Medicine and Biology**, v. 11, n. 1, p. 37-47, 2012. doi: 10.1007/s12522-011-0105-6.

154 MIEUSSET, R. *et al*. Association of scrotal hyperthermia with impaired spermatogenesis in infertile men. **Fertility and Sterility**, v. 48, n. 6, p. 1006-1011, 1987. doi: 10.1016/s0015-0282(16)59600-9.

155 CARLSEN, E. *et al*. History of febrile illness and variation in semen quality. **Human Reproduction**, v. 18, n. 10, p. 2089-2092, 2003. doi: 10.1093/humrep/deg412.

156 CHOWDHURY, A. K.; STEINBERGER, E. Early changes in the germinal epithelium of rat testes following exposure to heat. **Journal of Reproduction and Fertility**, v. 22, n. 2, p. 205-212, 1970. doi: 10.1530/jrf.0.0220205.

157 DURAIRAJANAYAGAM, D.; AGARWAL, A.; ONG, C. Causes, effects and molecular mechanisms of testicular heat stress. **Reproductive Biomedicine Online**, v. 30, n. 1, p. 14-27, 2015. doi: 10.1016/j.rbmo.2014.09.018.

158 DILUIGI, A. *et al*. Meiotic arrest in human oocytes is maintained by a Gs signaling pathway. **Biology of Reproduction**, v. 78, n. 4, p. 667-672, 2008. doi: 10.1095/biolreprod.107.066019.

159 TRIPATHI, A.; KUMAR, K. V. P.; CHAUBE, S. K. Meiotic cell cycle arrest in

mammalian oocytes. **Journal of Cellular Physiology**, v. 223, n. 3, p. 592-600, 2010. doi: 10.1002/jcp.22108.

160 GRINSTED, J. *et al*. Is low temperature of the follicular fluid prior to ovulation necessary for normal oocyte development? **Fertility and Sterility**, v. 43, n. 1, p. 34–39, 1985. doi: 10.1016/s0015-0282(16)48314-7.

161 MEZQUITA, B.; MEZQUITA, C.; MEZQUITA, J. Marked differences between avian and mammalian testicular cells in the heat shock induction and polyadenylation of Hsp70 and ubiquitin transcripts. **FEBS Letters**, v. 436, n. 3, p. 382-386, 1998. doi: 10.1016/s0014-5793(98)01172-7.

162 PADHI, A.; GHALY, M. M.; MA, L. Testis-enriched heat shock protein A2 (HSPA2): adaptive advantages of the birds with internal testes over the mammals with testicular descent. **Scientific Reports**, v. 6, n. 1, p. 1-7, 2016. doi: 10.1038/srep18770.

163 PABST, D. A.; ROMMEL, S. A.; MCLELLAN, W. A. Evolution of thermoregulatory function in cetacean reproductive systems. *In*: THEWISSEN, J. G. M. (ed.) **The emergence of whales:** evolutionary patterns in the origin of cetacea. Boston, MA: Springer, 1998. p. 379-87. doi: 10.1007/978-1-4899-0159-0_13.

164 ROMMEL, S. A.; PABST, D. A.; MCLELLAN, W. A. Reproductive thermoregulation in marine mammals: how do male cetaceans and seals keep their testes cool without a scrotum? it turns out to be the same mechanism that keeps the fetus cool in a pregnant female. **American Scientist**, v. 86, n. 5, p. 440-448, 1998.

165 DAVIS, R. W. **Marine mammals:** adaptations for an aquatic life. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-319-98280-9.

166 ROMMEL, S. A. *et al*. Anatomical evidence for a countercurrent heat exchanger associated with dolphin testes. **The Anatomical Record**, v. 232, n. 1, p. 150-156, 1992. doi: 10.1002/ar.1092320117.

167 NEWMAN, J. R. B. *et al*. Disease-specific biases in alternative splicing and tissue-specific dysregulation revealed by multitissue profiling of lymphocyte gene expression in type 1 diabetes. **Genome Research**, v. 27, n. 11, p. 1807-1815, 2017. doi: 10.1101/gr.217984.116.

ANNEX – Published works

Two articles were published during the master's degree:

- "A blueprint of septin expression in human tissues", abstract:

Septins are GTP-binding proteins that polymerize to form filaments involved in several important biological processes. In human, 13 distinct septins genes are classified in four groups. Filaments formed by septins are complex and usually involve members of each group in specific positions. Expression data from GTEx database, a publicly available expression database with thousands of samples derived from multiple human tissues, was used to evaluate the expression of septins. The brain is noticeably a hotspot for septin expression where few genes contribute to a large portion of septin transcript pool. Co-expression data between septins suggests two predominant specific complexes in brain tissues and one filament in other tissues. SEPT3 and SEPT5 are two genes highly expressed in the brain and with a strong co-expression in all brain tissues. Additional analysis shows that the expression of these two genes is highly variable between individuals, but significantly dependent on the individual's age. Age-dependent decrease of expression from those two septins involved in synapses reinforces their possible link with cognitive decay and neurodegenerative diseases associated with aging. Analysis of enrichment of Gene Ontology terms from lists of genes consistently co-expressed with septins suggests participation in diverse biological processes, pointing out some novel roles for septins. Interestingly, we observed strong consistency of some of these terms with experimentally described roles of septins. Coordination of septins expression with genes involved in DNA repair and cell cycle control may provide insights for previously described links between septins and cancer.

ZUVANOV, L.; MOTA, D. M. D.; ARAUJO, A. P. U.; DEMARCO, R. A blueprint of septin expression in human tissues. **Functional & Integrative Genomics**, v. 19, n. 5, p. 787-97, 2019. doi: 10.1007/s10142-019-00690-3.

- "Dissemination of *bla*$_{KPC-2}$ in an NTE$_{KPC}$ by an IncX5 plasmid", abstract:

bla$_{KPC-2}$ is disseminated worldwide usually in Tn*4401*, a Tn*3*-family transposon, and primarily in *Klebsiella pneumoniae* ST258, a well-known lineage that is distributed worldwide and responsible for several outbreaks. Although occurring rarely, *bla*$_{KPC-2}$ has been described in non-Tn*4401* elements (NTE$_{KPCs}$), first in China and then in a few other countries. This study reports the dissemination of a *bla*$_{KPC-2}$-carrying NTE$_{KPC}$ among ST11/CG258 *K. pneumoniae* strains and ST1642 *K. quasipneumoniae* subsp. *quasipneumoniae* AMKP9 in an Amazonian hospital. The dissemination was due to pAMKP10, an ~48 kbp IncX5 plasmid carrying ΔIS*Kpn6*/*bla*$_{KPC-2}$/IS*Kpn27* in a Tn*1722*-based unit. Although similar to NTE$_{KPC}$-Ia from pKP048

described in China, a different transposase is present upstream of IS*Kpn27*. Additionally, mutations were identified downstream of IS*Kpn27* but did not affect the *bla*$_{KPC-2}$ promoter regions. pAMKP10 conjugated *in vitro* only from CG258 isolates. Since CG258 strains are generally well adapted to the hospital environment, it is significant that pAMKP10 has found its way into this clinically significant clonal group. The impact of inter- and intraspecies dissemination of NTE$_{KPCs}$ and IncX5 plasmids harboring carbapenem resistance genes is unknown, but monitoring these plasmids could reveal their dissemination preferences.

SOUZA, R. C.; DABUL, A. N.; BORALLI, C. S.; **ZUVANOV, L.;** CAMARGO, I. L. B. C. Dissemination of blaKPC-2 in an NTEKPC by an IncX5 plasmid. **Plasmid**, v. 106, p. 102446, 2019. doi: 10.1016/j.plasmid.2019.102446.