Comprehensive Analysis of Retrotransposon Insertions within the Survival Motor Neuron Genes Involved in Spinal Muscular Atrophy

Albano Pinto
University of Tras-os-Montes and Alto Douro: Universidade de Tras-os-Montes e Alto Douro
Catarina Cunha
University of Tras-os-Montes and Alto Douro: Universidade de Tras-os-Montes e Alto Douro
Raquel Chaves
University of Tras-os-Montes and Alto Douro: Universidade de Tras-os-Montes e Alto Douro
Matthew ER Butchbach
Nemours Biomedical Research
Filomena Adega (filadega@utad.pt)
Universidade de Trás-os-Montes e Alto Douro https://orcid.org/0000-0001-5646-5534

Research Article

Keywords: spinal muscular atrophy, transposable elements, retrotransposons, genome dynamics, SMN1, SMN2

Posted Date: January 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1239736/v1

Abstract

Transposable elements (TEs) are interspersed repetitive DNA sequences with the ability to mobilize in the genome. The recent development of improved tools for evaluating TE-derived sequences in genomic studies has enabled an increasing attention to the contribution of TEs to human development and disease. Spinal muscular atrophy (SMA) is an autosomal recessive motor neuron disease that is caused by deletions or mutations in the *Survival Motor Neuron 1* (*SMN1*) gene. *SMN2* gene is a nearly perfect duplication of *SMN1*. Both genes (collectively known as *SMN1*/*SMN2*) are highly enriched in TEs. A comprehensive analysis of TEs insertions in the *SMN1/2 loci* of SMA carriers, patients and healthy/control individuals was completed to perceive TE dynamics in *SMN1/2* and try to establish a link between these elements and SMA.

We found an Alu insertion in the promoter region and one L1 element in the 3'UTR that likely play an important role as an alternative promoter and as an alternative terminator to the gene, respectively. Additionally, the several Alu repeats inserted in the genes' introns influence splicing, giving rise to alternative splicing events that cause RNA circularization and the birth of new alternative exons. These Alu repeats present throughout the genes are also prone to recombination events that can lead to *SMN1* exons deletions, that ultimately lead to SMA.

The many good and bad implications associated with the presence of TEs inside *SMN1/2* make this genomic region ideal for understanding the implications of TEs on genomic evolution as well as on human genomic disease.

Introduction

Transposable elements (TEs) are interspersed repetitive DNA sequences with the ability to mobilize in the genome. This mobility is mediated by element-encoded proteins such as DNA transposase or reverse transcriptase and occurs within the genome of virtually all walks of life, including prokaryotes, unicellular and multicellular eukaryotes and even large DNA viruses (Hancks and Kazazian Jr 2016; Jangam et al. 2017). TEs can be divided in two major classes based on their mechanism of transposition and each class can be subdivided into subclasses based on the mechanism of chromosomal integration (Bourque et al. 2018; Platt II et al. 2018; Cosby et al. 2019). Class 1 elements, also known as retrotransposons, have a "copy-and-paste" mobilization mechanism where an RNA intermediate is reverse-transcribed into a cDNA copy that is integrated elsewhere in the genome (Garcia-Perez et al. 2016; Bourque et al. 2018; Paço et al. 2015). Class 2 transposons mobilize via a DNA intermediate associated with a transposase, directly through a "cut-and-paste" mechanism or in the case of rolling-circle transposons, a "peel-and-paste" replicative mechanism involving a circular DNA intermediate (Bourque et al. 2018; Platt II et al. 2018). DNA transposons no longer mobilize in the human genome, with no evidence of their activity in our genome within the past 50 million years (Lander et al. 2001; Klein and O'Neill 2018).

Class 1 elements are divided in two major subclasses, long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. There are two types of non-LTR retrotransposons, long and short interspersed nuclear elements (LINEs and SINEs, respectively). Chromosomal integration of these elements is coupled to the reverse transcription via target-primed reverse transcription (Garcia-Perez et al. 2016; Bourque et al. 2018). LINE-1 (L1) elements contain two open reading frames (ORFs), an RNA binding and nuclear chaperone protein (ORF1) and a nuclease and reverse transcriptase protein (ORF2), that give these elements the ability to autonomously mobilize (Hancks and Kazazian Jr 2016; Platt II et al. 2018; Paço et al. 2015). A third short protein (ORF0) was recently described in primate L1 elements but its function remains unknown (Platt II et al. 2018). The only active autonomous TE family in humans is the LINE-1 family, although more than 99.9% of these L1 copies are fixed and no longer move due to various forms of rearrangements, mutations and 5'-truncation (Bourque et al. 2018; Payer and Burns 2019). The human genome still retains approximately 100 active L1 element copies capable of transposition, belonging to the L1PA1 and L1PA2 subfamilies (Bourque et al. 2018; Payer and Burns 2019; Burns 2020). The LINE-2 (L2) subfamily is present in the human genome but are no longer able to mobilize in the human genome, having lost this ability around 80 to 100 million years ago (Lander et al. 2001; Ayarpadikannan and Kim 2014).

SINEs structure is a combination of a 5' head that is derived from a ribosomal or tRNA pseudogene and a 3' tail homologous to a LINE (Platt II et al. 2018; Cosby et al. 2019). The LINE-like region of SINE elements is used to “parasitize” the enzymatic machinery of LINEs that allows for SINEs to mobilize (Platt II et al. 2018). Active SINEs, such as the primate-specific Alu elements, depend on the endonuclease and reverse transcriptase functions of the L1 elements for their genome propagation (Garcia-Perez et al. 2016; Bourque et al. 2018; Burns 2020). Within the human genome, Alu sequences can be divided into subfamilies of related elements based upon sequence variations shared by subfamily members, with the subfamilies being organized accordingly to their evolutionary age, with (J) standing for the oldest Alu subfamily, (S) for the intermediate and (Y) for the youngest subfamily (Batzer et al. 1996). Thus far, only AluYa5, AluYb8 and AluYb9 are the retrotransposition-competent Alu subfamilies in the modern human genome (Burns 2020).

Despite being largely inactive in the human genome, mammalian-wide interspersed repeats (MIRs) are the most ancient family of TEs (Jjingo et al. 2014). The SINE MIR was perfectly adapted for reverse transcription by L2, meaning that when L2 elements became fossils it also condemned MIR elements to the same fate (Lander et al. 2001; Ayarpadikannan and Kim 2014). Despite being inactive, such TE fossils often still contain coding or non-coding elements able to interact with the with the remaining genome, therefore still worth of being taken in consideration (Chuong et al. 2017). Mammalian genomes are dominated by satellite DNA elements (Chaves *et al.* 2017) and TEs, which can reach copy numbers in the hundreds of thousands and comprising from 45–50% of the human genome (Platt II et al. 2018; Burns 2020). Owing to their replicative nature and continuous accumulation, all TE families may sum up a total of 4.5 million copies in the human genome (Piégu et al. 2015; Platt II et al. 2018; Burns 2020; Nishihara 2019). In humans, the LINE class of fully autonomous retroelements represents approximately 17 to 21% of the genome with the non-autonomous SINE class of retrotransposons accounting for 11-

14% (Tam et al. 2019; Weber et al. 2019). Long terminal repeat retrotransposons represent 8% and DNA transposons account for approximately 3% of the human genome (Garcia-Perez et al. 2016; Weber et al. 2019; Tam et al. 2019).

TEs are highly abundant but only in the recent years their role in the genome started to be gradually recognized. Initially thought as inert remnants of evolution and mere genomic parasites, the advances in genome sequencing and the advent of more advanced tools gave us a new perspective and the recognition of TEs as important players in genomic evolution, genome organization and gene regulation. Once co-opted by the host genomes, TEs provide important sources of new regulatory sequences that can act as alternative promoters, tissue specific enhancers, splice sites, polyadenylation signals, insulators, termination sites and transcriptional factor binding sites thereby altering nearby gene expression in *cis* (Rebollo et al. 2012; Friedli and Trono 2015; Garcia-Perez et al. 2016; Chuong et al. 2017; Zeng et al. 2018; Sundaram and Wysocka 2020). Additionally, many TEs produce small RNAs—such as PIWI-interacting RNAs, small interfering RNAs, microRNAs, long noncoding RNAs or enhancer-overlapping RNAs—that are capable of altering transcription in *trans* (Friedli and Trono 2015; Chuong et al. 2017). TEs also can impact mammalian development and evolution through their domestication or the domestication of their proteins. In other words, TE transcripts perform functions for the host organism therefore increasing genetic diversity beyond that encoded by the germline genome (Garcia-Perez et al. 2016). Accordingly, the human genome contains around 50 to 150 genes that are probably domesticated or derived from TEs (Lander et al. 2001; O'Donnell and Burns 2010; Garcia-Perez et al. 2016). Exonization of intronic TEs—i.e. intronic TEs that are incorporated into exons of coding or noncoding transcripts—also gives these elements the ability to expand the mammalian transcriptome and proteome. Intronic Alu elements are particularly prone to be captured as alternative exons through cryptic splice sites residing within key positions of their sequences (Vorechovsky 2010; Friedli and Trono 2015; Garcia-Perez et al. 2016; Bourque et al. 2018; Zeng et al. 2018).

Despite the benefic role of TEs in genome evolution, their presence also holds disadvantages. TE activity can be detrimental to the host and cause several problems to normal gene expression and to genome organization, stability and integrity (Garcia-Perez et al. 2016; Chuong et al. 2017; Saleh et al. 2019; Payer and Burns 2019). The ability of TEs to transpose is the main mechanism associated with TE-induced diseases, with TE insertions into genes acting as insertional mutagens and interfering with gene function (O'Donnell and Burns 2010; Garcia-Perez et al. 2016; Saleh et al. 2019). Not surprisingly, the more actively propagating TE families in the human genome are the ones responsible for the onset of some of these diseases, namely (and in order of prevalence) Alu, L1 and SVA families (O'Donnell and Burns 2010). *De novo* germline and somatic TE insertions disrupting normal gene function have been implicated in several human diseases, among them neurologic disorders and cancer (Chuong et al. 2017; Saleh et al. 2019; Payer and Burns 2019; Burns 2020).

Alu elements and other TE families are also capable of promoting chromosomal rearrangements through its proneness to recombination events that may disrupt genomic stability (Hedges and Deininger 2007; Robberecht et al. 2013; Song et al. 2018; Lapp and Hunter 2019). These recombination events can occur

between the highly homologous regions dispersed by related TEs at distant genomic positions and result in small and large-scale deletions, duplications and inversions (Hedges and Deininger 2007; Bourque et al. 2018; Lapp and Hunter 2019). Another way TEs can pose a problem to the host genome stability and coding potential is through their ability to influence gene splicing. Alu elements—in particular inverted Alu repeats located within introns—are capable of influencing mRNA splicing, resulting in the formation of circular RNAs and altered splicing patterns (Payer and Burns 2019). This may result in nuclear retention of RNA and loss of protein-coding potential, making TEs potential agents of disease-causing events (Payer and Burns 2019; Percharde et al. 2020).

Attention to the contribution of TEs to neurodegenerative diseases has been rising in the last few years (Jönsson et al. 2020). Spinal muscular atrophy (SMA) is an early-onset, autosomal recessive neurological disease characterized by degeneration of motor neurons in the anterior horn of the spinal cord and brainstem nuclei (Crawford and Pardo 1996; Kolb and Kissel 2015). This motor neuron degeneration leads to progressive muscle weakness and atrophy. This neurodegenerative disease affects approximately 1/6,000 to 1/10,000 individuals and is the most common inherited cause of childhood mortality (Butchbach 2021). The carrier frequency for SMA is highly variable between populations, ranging between 1/25 to 1/50 (Wirth 2021; Butchbach 2021). Most cases of SMA result from a loss of *Survival Motor Neuron 1* (*SMN1*) (Lefebvre et al. 1995). This loss can be the result of *SMN1* deletion or conversion of *SMN1* into the paralogous *Survival Motor Neuron 2* (*SMN2*) gene (Wirth 2021; Butchbach 2021). The remaining 4% of SMA patients have subtle, intragenic mutations in *SMN1* (Butchbach 2021). Many of these subtle mutations are missense mutations that disturb the proper *SMN1* protein function (Wirth 2021; Butchbach 2021). There are 20 single nucleotide differences between *SMN1* and *SMN2*, with the C to T transition in exon 7 (c.840C>T) being the most functionally relevant difference (reviewed in (Butchbach 2016)). This substitution prompts an alternative splicing event on *SMN2* pre-mRNAs that leads to the skipping of exon 7 in the vast majority of *SMN2* transcripts (Lorson et al. 1999; Monani et al. 1999a). As a result, about 85% of the proteins produced by *SMN2* (SMNΔ7) are truncated and unstable, as the absence of exon 7 disrupts SMN protein oligomerization, making it unable to compensate for the lack of SMN production in the event of a loss of function by *SMN1* (Lorson and Androphy 2000; Cho and Dreyfuss 2010; Burnett et al. 2009).

SMN1 is located in a highly unstable region of the large arm of chromosome 5 (5q13.2), a region of the genome that is enriched in repeated sequences, pseudogenes and transposable elements (Campbell et al. 1997; Courseaux et al. 2003; Schmutz et al. 2004). It has been hypothesized that TEs may be the cause behind the instability of this region, given the ability of this mobile genetic elements to promote genetic instability and large chromosomal rearrangements observed in 5q13.2 (Campbell et al. 1997; Courseaux et al. 2003). *SMN1* and *SMN2* are both highly enriched in TEs—especially Alu and L1 repeats—spanning the whole gene (Ottesen and Singh 2020). The high abundance of Alu elements and other transposable elements in *SMN1* introns have an impact on the regulation of the splicing patterns as two Alu elements can give rise to new alternative *SMN* exons as well as on circularization events of *SMN* RNA that result from inverted Alu repeats (Ottesen et al. 2017; Pagliarini et al. 2020; Ottesen and Singh 2020). Such

accumulation of Alu repeats in *SMN1* introns also makes this gene prone to deletion events caused by Alu/Alu recombination events (Ruhno et al. 2019; Wirth et al. 1999; Jedličková et al. 2020).

The *SMN1* and *SMN2* (*SMN1/2*) genes are highly enriched in TEs and these mobile genetic elements may play a role in these genes regulation, in addition to being directly or indirectly involved in *SMN1* deletions. Prior studies, however, only used the reference gene sequence in their analysis. As several TE families are actively transposing and increasing in copy number within the human genome, most TE insertions are not present in the reference genome assembly. Analysis of the whole genome using next generation sequencing (NGS) technologies allows the detection of segregating structural variants within human populations and genotyping of transposable elements in healthy and diseased individuals. In this study, we compared the location and orientation of TE insertions in the *SMN1/2* loci of SMA patients, SMA carriers and healthy individuals. The results of this comprehensive analysis could provide important insights into the potential involvement of TEs in SMA onset as well as help understand the roles of TE dynamics in genome evolution, gene regulation and in human disease.

Results

Transposable elements and SMN1 transcription

As previously stated, *SMN1* and *SMN2* are highly enriched in TEs, including in key regions for gene transcription, such as the promoter and terminator regions. Our analysis of the *SMN1* promoter identified insertion of an AluJb repeat inside the promoter regions (**Figure 1**). This AluJb sequence harbors several transcription regulatory motifs upstream of the most used transcription start site (TSS1) including a fetal transcription start site (TSS2) and another transcription start site (TSS3) (**Figure 1**). The EMBOSS Nucleotide Analysis tool identified many other regulatory motifs within this Alu element sequence (**Figure 1**). Our analysis showed that all samples in study, SMA patients, carriers, non-carriers, healthy and the reference *SMN1/SMN2* sequence from Ensembl exhibit this Alu insertion in the promoter as well as harboring the two alternative transcriptional start sites and remaining regulatory motifs (TSS2 and TSS3) (**Figure 2**).

We analyzed in detail the gene promoter region and predicted transcriptional elements and other regulatory motifs. Interestingly, interaction between the cAMP response element-binding protein (CREB) located inside the AluJb sequence and the canonical cAMP response element (CRE) located between TSS1a and the start codon of *SMN1/2* has been previously demonstrated (Germain-Desprez et al. 2001; Monani et al. 1999b; Singh et al. 2017).

Regarding the gene region downstream of exon 7, our first analysis of *SMN1* gene reference sequence obtained from Ensembl showed several TE insertions belonging to various subfamilies (**Figure 3A**). The last canonical exon (exon 8) of *SMN1/2* is located within the terminator region of the gene and primarily serves as the 3'UTR region of the gene (Ottesen et al. 2017; Singh et al. 2020). We detected a L1 insertion within this exon. This insertion is a truncated L1 insertion corresponding to the 3' end of a L1MC5a subfamily retrotransposon (**Figure 3B**). Our analysis of the remaining samples in study showed that

independently of being SMA patients (exhibiting *SMN1* deletion previously, based on prior analysis), SMA carriers, non-carriers or healthy genomes, all show this L1 insertion inside exon 8 (**Figure 4**). Having in consideration that exon 8 serves as the 3'UTR region of the gene and the fact that there is a L1 insertion inside exon 8 in all samples in study, independently of disease status, we argue that the insertion of this retrotransposon in the 3'UTR region resulted in a domestication event that gave the gene an alternative terminator, here referred for the first time.

To determine whether L1MC5a element inserted in exon 8 is being expressed in *SMN1* coding sequence, we conducted a deeper analysis of the *SMN1* CCDS. Analysis of the longest *SMN1* transcript (GenBank: BC062723.1) in Dfam showed a L1 element at the 3'end of the CCDS, indicating that these complete *SMN1* transcripts have an imbedded TE sequence derived from the L1MC5a insertion in exon 8 (**Figure 5**). Interestingly, the predominant transcript of *SMN1*, isoform d (CCDS34181.1), does not contain this L1 insertion (**Figure 5**). These results suggest that the alternative terminator function of the L1 element in exon 8 is correct since the L1 insertion is only detected in the longest isoform of *SMN1* transcripts and effectively functions as an alternative transcription terminator to the canonical *SMN1* terminator.

Since *SMN1* and *SMN2* share extensive sequence homology and nearly identical TE insertional patterns, we analyzed *SMN2* transcripts and found that the longest isoform (BC000908.2) presents the same L1MC5a insertion in a similar 3' location to that seen in the longest *SMN1* transcript (results not shown). Similar to *SMN1*, the remaining and more common *SMN2* transcripts isoforms (d, a, b and c) do not exhibit any L1 insertion in their sequence (results not shown).

Transposable elements and alternative splicing by exonization

SMN1/2 introns are highly enriched in Alu-derived repeats with many of them in an inverted orientation (**Figure 6**). These inverted Alu repeats—mainly AluY, AluJr and AluSx1—span the whole *locus* but accumulate particularly in intron 1, 2a, 4 and 6. Our analysis of the remaining samples showed that all samples have the same pattern of Alu insertions in *SMN1/2* introns with the key Alu repeats involved in the genes' transcripts circularization being located in the same position and orientation as in the *SMN1* reference sequence obtained from Ensembl (**Figure 7**).

Another important role of these inverted Alu repeats located in *SMN1/2* introns is a complex TE domestication event that gives the genes new coding sequences, in a process termed exonization. Among the several TE families capable of forming new exons, Alu elements are particularly prone to be domesticated as alternative exons (Vorechovsky 2010; Bourque et al. 2018). It has been estimated that 5% of all alternatively spliced human exons derive from the exonization of Alu elements (Sorek et al. 2002; Piriyapongsa et al. 2007). In *SMN1* and *SMN2*, two alternative exons resulting from exonization have been reported thus far, exon 6B (Seo et al. 2016; Yoshimoto et al. 2016) and exon 9 (Ottesen et al. 2019). These two exonization events differ in the Alu subfamily involved, as well in the Alu arm involved

in the exonization. Exon 6B results from exonization of the left arm of an inverted AluY element and exon 9 originates from the right arm of an antisense AluSz element (**Figure 8**).

We identified some insertional polymorphisms regarding the Alu subfamily inserted in exon 6B and exon 9 gene locations. In the case of exon 6B, most samples showed an expected inverted AluY insertion but we also found some cases where the exonization involved an inverted AluSc8 insertion instead (**Figure 9**). Because of the high similarity of the *consensus* sequences of AluY and AluSc8 (98.4% sequence identity) and the fact that AluSc8 subfamily is thought to be the evolutionary progenitor of the younger (Y) Alu subfamily (Bao et al. 2015; Martinez-Gomez et al. 2020), we hypothesize that the two different results obtained is due to an incorrect prediction made by the algorithm as opposed to TE insertion variability in this location. Additionally, the AluSc8 insertion was only observed in sequences obtained from the 1000 Genomes Project database. In other words, this AluSc8 insertion may be due to low sequencing read depth that is characteristic of the 1000 Genomes Project samples.

Extensive insertional polymorphisms were also detected for exon 9 in the analyzed samples. Instead of the expected AluSz TE as described by (Ottesen et al. 2019), we observed AluSz6 and AluSx insertions in addition to the expected AluSz, as described (**Figure 10**). Since there is no clear connection between this polymorphic insertion and SMA disease state, we conclude that the polymorphism of this insertion may be due to interindividual variability.

Transposable elements and partial deletions of SMN1

The most direct link between activity of TEs and SMA onset stems from their ability to mediate recombination events that are known to lead to disease-associated deletions and other genomic rearrangements (Jahic et al. 2016; Payer and Burns 2019). The presence of several Alu repeats within *SMN1/2* that are in close proximity to each other, as previously discussed, make these genes particularly prone to Alu/Alu recombination events. The first Alu mediated deletion reported in *SMN1* is a deletion involving a large sequence of the gene from intron 4 to intron 6, involving exons 5 and 6 (Wirth et al. 1999). Our analysis of this breakpoint revealed that an AluSx1 element located in intron 4 and an AluSx3 in intron 6 are the Alu subfamilies responsible for the recombination event that led to exons 5 and 6 deletion (**Figure 11**; yellow box). Ruhno and colleagues (Ruhno et al. 2019) recently reported a partial deletion of the critical exons 7 and 8 whose breakpoints were within the Alu-rich intron 6 and the gene 3'UTR (Ruhno et al. 2019). Our analysis revealed an AluSx1 element in intron 6 directly upstream of exon 7-and an AluSx insertion in the 3'UTR region downstream of exon 8 (**Figure 11**; red box). An Alu/Alu recombination event was reported in a SMA patient with a deletion of *SMN1* exons 2A, 2B, 3, 4 and 5 (Jedličková et al. 2020). Our analysis of this deletion event showed that an antisense AluSp of intron 1 is the most likely element to be involved in the recombination event with the antisense AluSq of intron 5 (**Figure 11**; blue box).

Discussion

In this work we have found that the sequences of the *SMN1/2* genes are highly enriched in TE insertions, including in key gene regions. These insertions may have important effects on regulation, splicing, expression and overall stability of the genes.

TE insertions in the 5'UTR and promoter region of protein coding genes are common events with whole-genome analyses showing that up to 25% of human genes have TEs in their promoter and/or untranslated regions (van de Lagemaat et al. 2003; Jordan et al. 2003; Huh et al. 2008). SINEs in particular seem to be highly represented in these regions, owing to their higher affinity to G+C rich genome regions (Jordan et al. 2003; Thornburg et al. 2006). The presence of TEs in 5' regions of genes brings an evolutionary advantage for both the TE and for the host genome. For the TE, it represents an opportunity for translation as insertion in this open chromatin environment promotes its expression, and therefore its transposition (Wells and Feschotte 2020; Burns 2020). For the host genome, the presence of TEs in these critical regions is a potential source of novel regulatory sequences by fusing with/replacing a canonical gene promoter or alternatively serving as an alternative promoter either upstream or downstream of the canonical transcription start site (Faulkner et al. 2009; Branco and Chuong 2020; Marasca et al. 2020; Sundaram and Wysocka 2020).

This hints to a complete domestication and integration of this TE in the gene regulatory network. However, the significance of the remaining motifs inside the AluJb sequence remains to be investigated. Additionally, the TSSs present inside the Alu sequence are tissue-specific and/or developmental stage-specific TSSs, with TSS2 being used as a fetal transcription start site and TSS3 use still unknown (Singh et al. 2020). We argue that this AluJb is most likely serving as an alternative promoter or even as a tissue/developmental stage-specific promoter to *SMN1/2*. The fact that we found this Alu insertion in all samples in study regardless of disease status, further indicates a complete domestication of this element as an alternative promoter of the genes. Therefore, this insertion has led to an increased complex regulatory network capable of altering *SMN* expression, both in *cis* and *trans*. Promoter regions harboring Alu elements are subject to regulation in *trans* by long noncoding RNAs (lncRNAs) (Hu et al. 2016).

The majority of human genes use alternative polyadenylation sites that are embedded in TEs, suggesting that TEs can influence the 3' end processing of host gene transcripts (Lee et al. 2008; Cowley and Oakey 2013). L1 and other TE insertions are capable of interfering with endogenous *cis* regulatory elements present in 3'UTRs by introducing miRNA binding sites, promoting RNA editing and introducing polyadenylation signals (Roy-Engel et al. 2005; Faulkner et al. 2009; Sela et al. 2010; Garcia-Perez et al. 2016; Chen and Yang 2017; Chuong et al. 2017). Polyadenylation signals contained within retrotransposon sequences often lead to truncated or elongated 3'UTRs of full-length gene transcripts by providing an alternative terminator. As a result, these TE capabilities could repress transcription from the affected gene. This observation strengthens our hypothesis that the L1 present in *SMN1/2* 3'UTR region serves as an alternative terminator for the genes transcription by giving rise to longer transcripts.

Two factors may explain how this L1 element became fixed in this gene region. First, exon 8 has a lower percentage of G+C content (36.4%) when compared with the whole gene region (42.3%). The lower G+C

content may have favored an L1 insertion as these elements have a bias towards lower G+C regions of the genome (Deininger et al. 2003). Second, the fact that exon 8 serves as the 3'UTR may have facilitated insertion of the L1 element as pressure against TE insertions is often relaxed in these regions (Medstrand et al. 2002; Faulkner et al. 2009; Kabelitz and Bäurle 2015). We speculate, however, that such a large L1 insertion inside the gene coding region most likely had a large impact in the gene sequence and regulation and therefore should have been under negative selection pressure. Accordingly, L1 elements are especially underrepresented within genes, particularly those in the same transcriptional direction as the gene, because of their size and interference originated by retroelement regulatory motifs like polyadenylation signals (Medstrand et al. 2002; Hancks and Kazazian Jr 2016; Bourque et al. 2018). Thus, the insertion of this L1 element in the last exon of *SMN1/2*, that functions as the 3'UTR of the gene, is an example of a relatively rare event. Furthermore, the L1 insertion in exon 8 does not belong to the active L1 elements (also known as "hot L1 elements") of the human genome that are composed only of the L1PA1 and L1PA2 subfamilies (Bourque et al. 2018; Burns 2020). Therefore, this insertion is likely to be fixed in the human genome and not the result of a recent transposition event.

3'UTR retrotransposon insertions reduce mRNA expression (Faulkner et al. 2009; Hancks and Kazazian Jr 2016; Lavi and Carmel 2018). The presence of this L1MC5a element in the longer *SMN1* isoform transcripts could explain why these transcripts are less common than their shorter counterparts. Interestingly, weakly expressing genes were found to be rich in LINE insertions what can be explained by the ability of L1 elements to disrupt transcriptional elongation based on the presence of strong polyA signals in their sequences that possibly function as transcriptional terminators (Jjingo et al. 2011). We argue that the alternative terminator provided by the L1 element is used less frequently than the canonical gene terminator located in exon 7 and is probably only used in a tissue/time-specific manner. Accordingly, alternative UTRs that are often provided by TE insertions, can determine tissue-specific functions of mRNAs (Hughes 2006; Sela et al. 2007). 3'UTR retrotransposons insertional events, like the L1 insertion in the 3' region of *SMN1/2* reported in this study, are only moderately selected against and may provide a gradual mechanism of evolution by which retrotransposons alter the expression profile and influence crucial gene networks in the human genome (Piriyapongsa et al. 2007; Faulkner et al. 2009; Petri et al. 2019).

Additionally, it is possible that this L1MC5a element has the same transcription terminator function in *SMN2* as it shares the same TE insertional patterns in its sequence with *SMN1*. Accordingly, the longest *SMN2* transcript isoform presents this L1 insertion in its 3' region, in an identical sequence position as in *SMN1* longest transcript. This hints to this retrotransposon having the same alternative terminator role in *SMN2* transcription to that in *SMN1* transcription. We hypothesize that this alternative terminator role can have implications in SMA severity, which would require further analysis.

Circular RNAs (circRNAs) are a widely expressed class of non-colinear RNAs generated in a diverse set of eukaryotic organisms (Kristensen et al. 2019). Due to their lack of 5' and 3' termini, these RNAs are extremely stable meaning that even small levels of circRNAs may affect cellular metabolism by sequestering/sponging miRNAs, sequestration and trafficking of proteins, regulation of transcription and

generation of short RNA-binding proteins (Greene et al. 2017; Kristensen et al. 2019). CircRNAs are important regulators of cellular physiology and also potential biomarkers of disease onset or progression (Haque and Harries 2017; Kristensen et al. 2019). Circular RNAs have been associated with various human diseases, particularly cancer, diabetes mellitus, cardiovascular diseases, chronic inflammatory diseases and neurological disorders (Greene et al. 2017; Kristensen et al. 2019; Chen et al. 2020b). The most common way in which circRNAs are generated is through backsplicing in which the 5' splice site of a downstream exon is paired with the 3' splice site of an upstream exon (Ottesen et al. 2019; Voellenkle et al. 2019). One of the defining features of backsplicing events appears to be the RNA secondary structure formed by inverted short repeats, especially Alu elements, within intronic sequences upstream and downstream of the 3' and 5' splice sites (Ottesen et al. 2019). The existence of several inverted Alu repeats here detected throughout the *SMN1* reference sequence explains the high levels of circularization of *SMN1/2* transcripts as Alu repeats located in introns 4 and 5 are especially active in this process (Ottesen et al. 2019; Pagliarini et al. 2020; Singh et al. 2020). Pairing between the longest and highly Alu-enriched regions, introns 1 and 6, could be favored by several inverted Alu repeats thus potentially favoring backsplicing between exon 6 and exon 2A (Pagliarini et al. 2020). Additionally, generation of circRNAs with exons 2A, 2B, 3 and 4 requires pairing of the 5' splice site of exon 4 with the 3' splice site of exon 2A, which is made possible by the fact that intron 1 contains numerous Alu elements that are capable of pairing with the intron 4 Alu elements (Ottesen and Singh 2020).

The presence of these inverted repeats that are in a favorable position to pair with each othe, is often associated with alternative splicing events leading to circRNA biogenesis. Given that high levels of Alu inverted repeats in *SMN1/2* introns may explain how these genes generate several circular RNAs, we believe that circRNA biogenesis occurs in healthy individuals and SMA patients to the same extent. It is possible, however, that dysregulation of these Alu repeats in SMA-affected genomes may lead to an increased formation of circRNAs coded by *SMN1*. These higher levels of circRNAs formation and the widespread alternative circularization of *SMN1/2* pre-mRNA may have a still undiscovered role in SMA onset or may contribute to worse SMA phenotypes, owing to circRNAs ability to interfere with the coding capacity of human genes (Kristensen et al. 2019; Ottesen et al. 2019; Pagliarini et al. 2020). Additionally, circRNA formation in *SMN1/2* may function as a potential biomarker for the genes' overall transcriptional/splicing stability since higher circRNAs levels indicate aberrant RNA splicing events that may be linked to SMA. Due to their loop-structure, circRNAs are highly resistant to exonucleases and are detectable in peripheral blood which means that they may hold great potential as disease biomarkers, especially in neurologic and neurodegenerative conditions (Voellenkle et al. 2019; Chen et al. 2020b).

We have confirmed two exonization events within the *SMN* genes, exon 6B (Seo et al. 2016; Yoshimoto et al. 2016) and exon 9 (Ottesen et al. 2019), that have been generated by different Alu insertions. The fact that in *SMN1* all known Alu recombination events that led to deletions occurred among Alu elements of the (S) subfamily, supports the idea that alongside proximity, sequence identity between the two elements at a locus also appears to be proportional to their chances of successful recombination (Sen et al. 2006; Jahic et al. 2016; Kim et al. 2016). The high conservation of position and orientation of the Alu insertions involved in these exonization events implies that these domestication events occur identically in healthy

and diseased genomes. Furthermore, they are important sources of novel exons that increase the coding capacity of the genes beyond the coding capacity of *SMN1/2* canonical exons. In a wider spectrum, Alu and other TEs provide transcriptome diversity and ultimately result in the diversification of the human proteome.

Some insertional polymorphisms that were observed in these Alu repeats between the analyzed samples may be the result of normal interpersonal sequence variability. Accordingly, polymorphic Alu elements account for 17% of structural variants in the human genome, clearly establishing a link between individual TE polymorphisms and human genetic variation (Kryatova et al. 2017). Since the SMA samples here analyzed are from individuals lacking the *SMN1* gene, we cannot draw definitive conclusions about the involvement of Alu elements in the deletion events in a disease context. Our results show that in the analyzed SMA patients and in the remaining samples, the critical Alu elements responsible for these deletions are also present in the same position and orientation in the *SMN1/2* sequence as in the gene reference sequence and as described previously (Ruhno et al. 2019; Wirth et al. 1999; Jedličková et al. 2020). Although these Alu insertions sites are most probably the reason for the complex deletion events, the high conservation of orientation and of subfamily type found in all the samples analyzed (healthy and diseased) implies that their presence *per se* is not the reason for the deletion events. Their presence in *SMN1/2* introns is a source of sequence homology that can be responsible for genomic rearrangements and consequently disease in some genomes (Jahic et al. 2016; Payer and Burns 2019). TE recombination may be responsible for the approximately 2% of SMA cases that result from *de novo* mutations in *SMN1* and are not inherited from carrier parents (Wirth 2021). It is likely that the several Alu repeats present throughout *SMN1* are responsible for *de novo* deletions in germinative cells due to unequal recombination since TE silencing mechanisms are often relaxed in these developmental stages (Bourque et al. 2018; Saleh et al. 2019; Burns 2020). Alu elements may indeed, play an important role in the high instability of the *SMN1/2* genomic region leading to disease-causing deletions of *SMN1* exons and potentially whole gene deletions under specific circumstances. Alu elements invasion of *SMN1/2* made these genes very susceptible to Alu-mediated deletions, that have critical consequences to genome stability and host health.

Conclusions

Our analysis of the *SMN* revealed a pervasive invasion of its sequence by TEs that we believe may severally impact these genes regulation structure, expression and overall genomic stability. The several TEs present inside these genes, especially Alu and L1 elements that are highly enriched in the promoter and intronic regions of the gene, seem to play important roles in gene expression, novel exon creation, alternative splicing and deletion events known to lead to SMA. Additionally, a L1 element insertion in the 3'UTR region of the gene is also responsible for a domestication event that gave the gene an alternative terminator, therefore increasing the diversity of *SMN1* transcripts and being a prime example of how a TE insertion inside a protein-coding gene can create a gradual mechanism of evolution by which retrotransposons alter the human transcriptome.

Materials And Methods

An extensive analysis of the *SMN1* and *SMN2* gene sequences was conducted on a comprehensive *SMN1/2* gene sequences cohort. Our sample database consisted of *SMN1/2* gene sequences from 20 spinal muscular atrophy (SMA) carriers, 20 non-carriers, 37 SMA patients, two healthy individuals and the reference *SMN1* gene sequence available in Ensembl (gene ID: ENSG00000172062, human genome assembly GRCh38.p13). Genome sequencing data of the carriers and non-carriers for SMA was downloaded from “1000 Genomes Project” online database (Consortium et al. 2010) and genome sequencing data from the 37 SMA patients and the two healthy individuals has been previously published by the Motor Neuron Diseases Research Laboratory (MNDRL) in collaboration with Illumina, Inc (Chen et al. 2020a). The SMA status for the MNDRL cohort was confirmed by digital droplet PCR (Stabley et al. 2015; Jiang et al. 2020; Stabley et al. 2021).

The raw next generation sequencing (NGS) data obtained was processed and then mapped against *SMN1* reference sequence by Geneious Mapper tool (Kearse et al. 2012) (Geneious Prime version 2020.0.5 software; https://www.geneious.com). The mapping sensitivity was set to medium and the number of iterations was set to 10 times. Geneious Mapper generated a contig of the multiple reads mapped to *SMN1* and a *consensus* sequence of the mapped reads which was used for predicting TE insertions.

All *SMN1/2* sequences obtained and the *SMN1* and *SMN2* Consensus Coding Sequences (CCDS) (Pujar et al. 2018) isoforms in study were screened for TEs insertions using Dfam version 3.1 (Hubley et al. 2016; Storer et al. 2021). All the TE annotated *SMN1/2* sequences and CCDS were aligned using Clustal Omega version 1.2.2 multiple sequence alignment (MSA) program (Sievers et al. 2011) (Geneious Prime version 2020.0.5). *SMN1/2* transcriptional elements and motifs, including the promoter elements and other regulatory sequences were either described previously (Germain-Desprez et al. 2001; Monani et al. 1999b) or were computationally predicted by EMBOSS Nucleotide Analysis version 1.1.1 (Rice et al. 2000) (Geneious Prime version 2020.0.5).

Abbreviations

circRNAs - circular RNAs

LINEs - Long Interspersed Nuclear Elements

lncRNAs - long noncoding RNAs

LTR - Long Terminal Repeat

MIR - Mammalian-wide Interspersed Repeats

NGS - Next Generation Sequencing

ORFs - Open Reading Frames

SINEs - Short Interspersed Nuclear Elements

SMA - Spinal Muscular Atrophy

SMN1 - Survival Motor Neuron 1

SMN2 - Survival Motor Neuron 2

TEs - Transposable elements

Declarations

Ethics Approval and Consent to Participate

This study involving human participants was performed in accordance with the ethical standards of the institutional and/or research committees (Nemours Institutional Review Board, #764456) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This study is registered with the ClinicalTrials.gov resource (NCT0175441 and NCT0253244).

Consent for Publication

The samples were de-identified so that no identifiable protected health information on the participants was known.

Availability of Data and Materials

The datasets analyzed during this study are available from Dr. Butchbach (Matthew.Butchbach@nemours.org) on reasonable request.

Competing Interests

The authors declare that they have no competing interests, either financial or non-financial.

Funding

This work was supported by the BioISI project UIDP/04046/2020 Research Unit grant from the Fundação para a Ciência e Tecnologia (FCT), Portugal (FA), Foundation for the National Institutes of Health, USA (MERB; P20GM103446 and P20GM103464) and the Nemours Foundation (MERB).

Authors' Contributions

AP performed the analyses and wrote the manuscript, CC assisted in the analyses and manuscript writing, RC assisted in data interpretation and manuscript revision, MERB provided the whole genome sequencing datasets and assisted in data interpretation as well as in manuscript writing and revision and

FA conceptualized the study, supervised implementation and analyses, assisted in data interpretation and manuscript writing and finalization. All authors read and approved the final manuscript.

Acknowledgements

We deeply thank Drs. Xiao Chen, Michael Eberle and Ryan Taft at Illumina Inc. for providing the genome sequencing dataset from 37 SMA patients and the two healthy individuals.

References

Ayarpadikannan, S and Kim, HS (2014). The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics and Informatics* 12: 98-104.

Bao, W, Kojima, KK and Kohany, O (2015). Repbase Update, a database of repetitive elements in eukaryotics genomes. *Mobile DNA* 6: 11.

Batzer, MA, Deininger, PL, Hellmann-Blumberg, U, Jurka, J, Labuda, D, Rubin, CM, Schmid, CW, Ziętkiewicz, E and Zuckerkandl, E (1996). Standardized nomenclature for Alu repeats. *Journal of Molecular Evolution* 42: 3-6.

Bourque, G, Burns, KH, Gehring, M, Gorbunova, V, Seluanov, A, Hammell, M, Imbeault, M, Izsvák, Z, Levin, HL, Macfarlan, TS, Mager, DL and Feschotte, C (2018). Ten things you should know about transposable elements. *Genome Biology* 19: 199.

Branco, MR and Chuong, EB (2020). Crossroads between transposons and gene regulation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 375: 20190330.

Burnett, BG, Muñoz, E, Tandon, A, Kwon, DY, Sumner, CJ and Fischbeck, KH (2009). Regulation of SMN protein stability. *Molecular and Cellular Biology* 29: 1107-1115.

Burns, KH (2020). Our conflict with transposable elements and its implications for human disease. *Annual Review of Pathology: Mechanisms of Disease* 15: 51-70.

Butchbach, MER (2016). Copy number variations in the *Survival Motor Neuron* genes: implications for spinal muscular atrophy and other neurodegenerative diseases. *Frontiers in Molecular Biosciences* 3: 7.

Butchbach, MER (2021). Genomic variability in the Survival Motor Neuron genes (*SMN1* and *SMN2*): implications for spinal muscular atrophy phenotype and therapeutics development. *International Journal of Molecular Sciences* 22: 7896.

Campbell, L, Potter, A, Ignatius, J, Dubowitz, V and Davies, K (1997). Genomic variation and gene conversion in spinal muscular atrophy: implications for disease process and clinical phenotype. *American Journal of Human Genetics* 61: 40-50.

Chaves, Raquel, Ferreira, Daniela, Mendes-da-Silva, Ana, Meles, Susana and Adega, Filomena (2017). FA-SAT Is an Old Satellite DNA Frozen in Several Bilateria Genomes. *Genome Biology and Evolution* 9(11): 3073-3087.

Chen, LL and Yang, L (2017). *ALU*ternative regulation for gene expression. *Trends in Cell Biology* 27: 480-490.

Chen, X, Sanchis-Juan, A, French, CE, Connell, AJ, Delon, I, Kingsbury, Z, Chawla, A, Halpern, AL, Taft, RJ, BioResource, NIHR, Bentley, DR, Butchbach, MER, Raymond, FL and Eberle, MA (2020a). Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genetics in Medicine* 22: 945-953.

Chen, YJ, Chen, CY, Mai, TL, Chuang, CF, Chen, YC, Gupta, SK, Yen, L, Wang, YD and Chuang, TJ (2020b). Genome-wide, integrative analysis of circular RNA dysregulation and the corresponding circular RNA-microRNA-mRNA regulatory axis in autism. *Genome Research* 30: 375-391.

Cho, S and Dreyfuss, G (2010). A degron created by SMN2 exon 7 skipping is a principal contributor to spinal muscular atrophy severity. *Genes and Development* 24: 438-442.

Chuong, EB, Elde, NC and Feschotte, C (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* 18: 71-86.

Consortium, 1000 Genomes Project, Abescasis, GR, Altshuler, D, Auton, A, Brooks, KD, Durbin, RM, Gibbs, RA, Hurles, ME and McVean, GA (2010). A map of human genome variation from population-scale sequencing. *Nature* 467: 1061-1073.

Cosby, RL, Chang, NC and Feschotte, C (2019). Host-transposon interactions: conflict, cooperation and cooption. *Genes and Development* 33: 1098-1116.

Courseaux, A, Richard, F, Grosgeorge, J, Ortola, C, Viale, A, Turc-Carel, C, Dutrillaux, B, Gaudray, P and Nahon, JL (2003). Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. *Genome Research* 13: 369-381.

Cowley, M and Oakey, RJ (2013). Transposable elements: re-wire and fine-tune the transcriptome. *Plos Genetics* 9: e1003234.

Crawford, TO and Pardo, CA (1996). The neurobiology of childhood spinal muscular atrophy. *Neurobiology of Disease* 3: 97-110.

Deininger, PL, Moran, JV, Batzer, MA and Kazazian Jr, HH (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics and Development* 13: 651-658.

Faulkner, GJ, Kimura, Y, Daub, CO, Wani, S, Plessy, C, Irvine, KM, Schroder, K, Cloonan, N, Steptoe, AL, Lassmann, T, Waki, K, Hornig, N, Arakawa, T, Takahashi, H, Kawai, J, Forrest, ARR, Suzuki, H, Hayashizaki,

Y, Hume, DA, Orlando, V, Grimmond, SM and Carninci, P (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* 41: 563-571.

Friedli, M and Trono, D (2015). The developmental control of transposable elements and the evolution of higher species. *Annual Review of Cell and Developmental Biology* 31: 429-451.

Garcia-Perez, JL, Widmann, TJ and Adams, IR (2016). The impact of transposable elements on mammalian development. *Development* 143: 4101-4114.

Germain-Desprez, D, Brun, T, Rochette, C, Semionov, A, Rouget, R and Simard, LR (2001). The SMN genes are subject to transcriptional regulation during cellular differentiation. *Gene* 279: 109-117.

Greene, J, Baird, AM, Brady, L, Lim, M, Gray, SG, McDermott, R and Finn, SP (2017). Circular RNAs: biogenesis, function and role in human diseases. *Frontiers in Molecular Biosciences* 4: 38.

Hancks, DC and Kazazian Jr, HH (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA* 7: 9.

Haque, S and Harries, LW (2017). Circular RNAs (circRNAs) in health and disease. *Genes* 8: 353.

Hedges, DJ and Deininger, PL (2007). Inviting instability: transposable elements, double-strand breaks and the maintenance of genome integrity. *Mutation Research* 616: 46-59.

Hu, S, Wang, X and Shan, G (2016). Insertion of an Alu element in a lncRNA leads to primate-specific modulation of alternative splicing. *Nature Structural & Molecular Biology* 23: 1011-1019.

Hubley, R, Finn, RD, Clements, J, Eddy, SR, Jones, TA, Bao, W, Smit, AFA and Wheeler, TJ (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research* 44: D81-D89.

Hughes, TA (2006). Regulation of gene expression by alternative untranslated regions. *Trends in Genetics* 22: 119-122.

Huh, JW, Kim, DS, Ha, HS, Lee, JR, Kim, YJ, Ahn, K, Lee, SR, Chang, KT and Kim, HS (2008). Cooperative exonization of MaLR and AluJo elements contributed an alternative promoter and novel splice variants of RNF19. *Gene* 424: 63-70.

Jahic, A, Erichsen, AK, Deufel, T, Tallaksen, CM and Beetz, C (2016). A polymorphic *Alu* insertion that mediates distinct disease-associated deletions. *European Journal of Human Genetics* 24: 1371-1374.

Jangam, D, Feschotte, C and Betrán, E (2017). Transposable element domestication as an adaptation to evolutionary conflicts. *Trends in Genetics* 33: 817-831.

Jedličková, I, Přistoupilová, A, Nosková, L, Majer, F, Stránecký, V, Hartmannová, H, Hodaňová, K, Trešlová, H, Hýblová, M, Solár, P, Minárik, G, Giertlová, M and Kmoch, S (2020). Spinal muscular atrophy caused by

a novel *Alu*-mediated deletion of exons 2a-5 in *SMN1* undetectable with routine genetic testing. *Molecular Genetics and Genomic Medicine* 8: e1238.

Jiang, L, Lin, R, Gallagher, S, Zayac, A, Butchbach, MER and Hung, P (2020). Development and validation of a 4-color multiplexing spinal muscular atrophy (SMA) genotyping assay on a novel integrated digital PCR instrument. *Scientific Reports* 10: 19892.

Jjingo, D, Conley, AB, Wang, J, Mariño-Ramírez, L and Jordan, IK (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA* 5: 14.

Jjingo, D, Huda, A, Gundapuneni, M, Mariño-Ramírez, L and Jordan, IK (2011). Effect of transposable element environment of human genes on gene length and expression. *Genome Biology and Evolution* 3: 259-271.

Jönsson, ME, Garza, R, Johansson, PA and Jakobsson, J (2020). Transposable elements: a common feature of neurodevelopmental and neurodegenerative disorders. *Trends in Genetics* 36: 610-623.

Jordan, IK, Rogozin, IB, Glazko, GV and Koonin, EV (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* 19: 68-72.

Kabelitz, T and Bäurle, I (2015). Get the jump--Do 3'UTRs protect transposable elements from silencing? *Mobile Genetic Elements* 5: 51-54.

Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper, A, Markowitz, S, Duran, C, Thierer, T, Ashton, B, Meintjes, P and Drummond, A (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647-1649.

Kim, S, Cho, CS, Han, KS and Lee, J (2016). Structural variation of *Alu* element and human disease. *Genomics and Informatics* 14: 70-77.

Klein, SJ and O'Neill, RJ (2018). Transposable elements: genome innovation, chromosome diversity and centromere conflict. *Chromosome Research* 26: 5-23.

Kolb, SJ and Kissel, JT (2015). Spinal muscular atrophy. *Neurologic Clinics* 33: 831-846.

Kristensen, LS, Andersen, MA, Stagsted, LVW, Ebbesen, KK, Hansen, TB and Kjems, J (2019). The biogenesis, biology and characterization of circular RNAs. *Nature Reviews Genetics* 20: 675-691.

Kryatova, MS, Steranka, JP, Burns, KH and Payer, LM (2017). Insertion and deletion polymorphisms of the ancient *AluS* family in the human genome. *Mobile DNA* 8: 6.

Lander, ES, Linton, LM, Birren, B , Nusbaum, C , Zody, MC , Baldwin, J, Devon, K, Dewar, K, Doyle, M, FitzHugh, W, Funke, R, Gage, D, Harris, K, Heaford, A, Howland, J, Kann, L , Lehoczky, J, LeVine, R, McEwan,

P, McKernan, K, Meldrim, J, Mesirov, JP, Miranda, C, Morris, W, Naylor, J*, et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Lapp, HE and Hunter, RG (2019). Early life exposures, neurodevelopmental disorders and transposable elements. *Neurobiology of Stress* 11: 100174.

Lavi, E and Carmel, L (2018). Alu exaptation enriches the human transcriptome by introducing new gene ends. *RNA Biology* 15: 715-725.

Lee, JY, Ji, Z and Tian, B (2008). Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Research* 36: 5581-5590.

Lefebvre, S, Bürglen, L, Reboullet, S, Clermont, O, Burlet, P, Viollet, L, Benichou, B, Cruaud, C, Millasseau, P, Zeviani, M, Le Paslier, D, Frézal, J, Cohen, D, Weissenbach, J, Munnich, A and Melki, J (1995). Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 80: 155-165.

Lorson, CL and Androphy, EJ (2000). An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Human Molecular Genetics* 9: 259-265.

Lorson, CL, Hahnen, E, Androphy, EJ and Wirth, B (1999). A single nucleotide in the *SMN* gene regulates splicing and is responsible for spinal muscular atrophy. *Proceedings of the National Academy of Sciences of the United States of America* 96: 6307-6311.

Marasca, F, Gasparotto, E, Polimeni, B, Vadalá, R, Ranzani, V and Bodega, B (2020). The sophicated transcriptional response governed by transposable elements in human health and disease. *International Journal of Molecular Sciences* 21: 3201.

Martinez-Gomez, L, Abascal, F, Jungreis, I, Pozo, F, Kellis, M, Mudge, JM and Tress, ML (2020). Few SINEs of life: Alu elements have little evidence for biological relevance despite elevated translation. *NAR Genomics and Bioinformatics* 2: lqz023.

Medstrand, P, van de Lagemaat, LN and Mager, DL (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Research* 12: 1483-1495.

Monani, UR, Lorson, CL, Parsons, DW, Prior, TW, Androphy, EJ, Burghes, AHM and McPherson, JD (1999a). A single nucleotide difference that alters splicing patterns distinguishes the SMA gene *SMN1* from the copy gene *SMN2*. *Human Molecular Genetics* 8: 1177-1183.

Monani, UR, McPherson, JD and Burghes, AHM (1999b). Promoter analysis of the human centromeric and telometic survival motor neuron genes (SMN^c and SMN^t). *Biochimica et Biophysica Acta* 1445: 330-336.

Nishihara, H (2019). Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes and Genetic Systems* 94:

269-281.

O'Donnell, KA and Burns, KH (2010). Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mobile DNA* 1: 21.

Ottesen, EW, Luo, D, Seo, J, Singh, NN and Singh, RN (2019). Human *Survival Motor Neuron* genes generate a vast repertoire of circular RNAs. *Nucleic Acids Research* 47: 2884-2905.

Ottesen, EW, Seo, J, Singh, NN and Singh, RN (2017). A multilayered control of the human *Survival Motor Neuron* gene expression by Alu elements. *Frontiers in Microbiology* 8: 2252.

Ottesen, EW and Singh, RN (2020). Characteristics of circular RNAs generated by human *Survival Motor Neuron* genes. *Cellular Signalling* 73: 109696.

Paço, Ana, Adega, Filomena and Chaves, Raquel (2015). LINE-1 retrotransposons: from 'parasite' sequences to functional elements. *Journal of Applied Genetics* 56(1): 133-145.

Pagliarini, V, Jolly, A, Bielli, P, Di Rosa, V, De la Grange, P and Sette, C (2020). Sam68 binds *Alu*-rich introns in *SMN* and promotes pre-mRNA circularization. *Nucleic Acids Research* 48: 633-645.

Payer, LM and Burns, KH (2019). Transposable elements in human genetic disease. *Nature Reviews Genetics* 20: 760-772.

Percharde, M, Sultana, T and Ramalho Santos, M (2020). What doesn't kill you makes you stronger: transposons as dual players in chromatin regulation and genomic variation. *Bioessays* 42: 1900232.

Petri, R, Brattas, PL, Sharma, Y, Jönsson, ME, Pircs, K, Bengzon, J and Jakobsson, J (2019). LINE-2 transposable elements are a source of functional human microRNAs and target sites. *Plos Genetics* 15: e1008036.

Piégu, B, Bire, S, Arensburger, P and Bigot, Y (2015). A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. *Molecular Phylogenetics and Evolution* 86: 90-109.

Piriyapongsa, J, Rutledge, MT, Patel, S, Borodovsky, M and Jordan, IK (2007). Evaluating the protein coding potential of exonized transposable element sequences. *Biology Direct* 2: 31.

Platt II, RN, Vangewege, MW and Ray, DA (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* 26: 25-43.

Pujar, S, O'Leary, NA, Farrell, CM, Loveland, JE, Mudge, JM, Wallin, C, Girón, CG, Diekhans, M, Barnes, I, Bennett, R, Berry, AE, Cox, E, Davidson, C, Goldfarg, T, Gonzalez, jM, Hunt, T, Jackson, JD, Joardar, V, Kay, MP, Kodali, VK, Martin, FJ, McAndrews, M, McGarvey, KM, Murphy, M, Rajput, B*, et al.* (2018). Consensus

coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research* 46: D221-D228.

Rebollo, R, Farivar, S and Mager, DL (2012). C-GATE – catalogue of gene affected by transposable elements. *Mobile DNA* 3: 9.

Rice, P, Longden, I and Bleasby, A (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-277.

Robberecht, C, Voet, T, Zamani Esteki, M, Nowakowska, BA and Vermeesch, JR (2013). Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Research* 23: 411-418.

Roy-Engel, AM, El-Sawy, M, Farooq, L, Odom, GL, Perepelitsa-Belancio, V, Bruch, H, Oyeniran, OO and Deininger, PL (2005). Human retroelements may introduce intragenic polyadenylation signals. *Cytogenetic and Genome Research* 110: 365-371.

Ruhno, C, McGovern, VL, Avenarius, MR, Snyder, PJ, Prior, TW, Nery, FC, Muhtaseb, A, Roggenbuck, JS, Kissel, JT, Sansone, VA, Siranosian, JJ, Johnstone, AJ, Nwe, PH, Zhang, RZ, Swoboda, KJ and Burghes, AHM (2019). Complete sequencing of the *SMN2* gene in SMA patients detect *SMN* gene deletion junctions and variants in SMN that modify the SMA phenotype. *Human Genetics* 138: 241-256.

Saleh, A, Marcia, A and Muotri, AR (2019). Transposable lements, inflammation and neurological disease. *Frontiers in Neurology* 10: 894.

Schmutz, J, Martin, J, Terry, A, Couronne, O, Grimwood, J, Lowry, S, Gordon, LA, Scott, D, Xie, G, Huang, W, Hellsten, U, Tran-Gyamfi, M, She, X, Prabhakar, S, Aerts, A, Altherr, M, Bajorek, E, Black, S, Branscomb, E, Caoile, C, Challacombe, JF, Chan, YM, Denys, M, Detter, JC, Escobar, J*, et al.* (2004). The DNA sequence and comparative analysis of human chromosome 5. *Nature* 431: 268-274.

Sela, N, Kim, E and Ast, G (2010). The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biology* 11: R59.

Sela, N, Mersch, B, Gal-Mark, N, Lev-Maor, G, Hotz-Wagenblatt, A and Ast, G (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biology* 8: R127.

Sen, SK, Han, KS, Wang, J, Lee, J, Wang, H, Callinan, PA, Dyer, M, Cordaux, R, Liang, P and Batzer, MA (2006). Human genomic deletions mediated by recombination between *Alu* elements. *American Journal of Human Genetics* 79: 41-53.

Seo, J, Singh, NN, Ottesen, EW, Lee, BM and Singh, RN (2016). A novel human-specific splice isoform alters the critical C-terminus of Survival Motor Neuron protein. *Scientific Reports* 6: 30778.

Sievers, F, Wilm, A, Dineen, D, Gibson, TJ, Karplus, K, Li, W, Lopez, R, McWilliam, H, Remmert, M, Söding, J, Thompson, JD and Higgins, DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539.

Singh, NN, Ottesen, EW and Singh, RN (2020). A survey of transcripts generated by spinal muscular atrophy genes. *Biochimica et Biophysica Acta* 1863: 194562.

Singh, RN, Howell, MD, Ottesen, EW and Singh, NN (2017). Diverse role of survival motor neuron protein. *Biochimica et Biophysica Acta* 1860: 299-315.

Song, X, Beck, CR, Du, R, Campbell, IM, Coban Akdemir, Z, Gu, S, Breman, AM, Stankiewicz, P, Ira, G, Shaw, CA and Lupski, JR (2018). Predicting human genes susceptible to genomic instability associated with *Alu*/*Alu*-mediated rearrangements. *Genome Research* 28: 1228-1242.

Sorek, R, Ast, G and Graur, D (2002). *Alu*-containing exons are alternatively spliced. *Genome Research* 12: 1060-1067.

Stabley, DL, Harris, AW, Holbrook, J, Chubbs, NJ, Lozo, KW, Crawford, TO, Swoboda, KJ, Funanage, VL, Wang, W, Mackenzie, W, Scavina, M, Sol-Church, K and Butchbach, MER (2015). *SMN1* and *SMN2* copy numbers in cell lines derived from patients with spinal muscular atrophy as measured by array digital PCR. *Molecular Genetics and Genomic Medicine* 3: 248-257.

Stabley, DL, Holbrook, J, Scavina, M, Crawford, TO, Swoboda, KJ, Robbins, KM and Butchbach, MER (2021). Detection of *SMN1* and *SMN2* gene conversion events and partial *SMN1* deletions using array digital PCR. *Neurogenetics* 22: 53-64.

Storer, J, Hubley, R, ROsen, J, Wheeler, TJ and Smit, AF (2021). The Dfam community resource of transposable element families, sequence models and genome annotations. *Mobile DNA* 12: 2.

Sundaram, V and Wysocka, J (2020). Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 375: 20190347.

Tam, OH, Ostrow, LW and Gale Hammell, M (2019). Diseases of the nERVous system: retrotransposon activity in neurodegenerative disease. *Mobile DNA* 10: 32.

Thornburg, BG, Gotea, V and MakaLowski, W (2006). Transposable elements as a significant source of transcription regulating signals. *Gene* 365: 104-110.

van de Lagemaat, LN, Landry, JR, Mager, DL and Medstrand, P (2003). Transposable elements in mammals promote regulatory variation ande diversification of genes with specialized functions. *Trends in Genetics* 19: 530-536.

Voellenkle, C, Perfetti, A, Carrara, M, Fuschi, P, Renna, LV, Longo, M, Sain, SB, Cardani, R, Valaperta, R, Silverstri, G, Legnini, I, Bozzoni, I, Furling, D, Gaetano, C, Falcone, G, Meola, G and Martelli, F (2019). Dysregulation of circular RNAs in myotonic dystrophy type 1. *International Journal of Molecular Sciences* 20: 1938.

Vorechovsky, I (2010). Transposable elements in disease-associated cryptic exons. *Human Genetics* 127: 135-154.

Weber, P, Rausch, C, Scholl, A and Cardoso, MC (2019). Repli-FISH (fluorescence in situ hybridization): application of 3D-(immuno)-FISH for the study of DNA replication timing of genetic repeat elements. *OBM Genetics* 3: 1901062.

Wells, JN and Feschotte, C (2020). A field guide to eukaryotic transposable elements. *Annual Review of Genetics* 54: 539-561.

Wirth, B (2021). Spinal muscular atrophy: in the challenge lies a solution. *Trends in Neurosciences* 44: 306-322.

Wirth, B, Herz, M, Wetter, A, Moskau, S, Hahnen, E, Rudnik-Schöneborn, S, Wienker, T and Zerres, K (1999). Quantitative analysis of survival motor neuron copies: identification of subtle *SMN1* mutations in patients with spinal muscular atrophy, genotype-phenotype correlation and implications for genetic counseling. *American Journal of Human Genetics* 64: 1340-1356.

Yoshimoto, S, Harahap, NIF, Hamamura, Y, Ar Rochmah, M, Shima, A, Morisada, N, Shinohara, M, Saito, T, Saito, K, Lai, PS, Matsuo, M, Awano, H, Morioka, I, Iijima, K and Nishio, H (2016). Alternative splicing of a cryptic exon embedded in intron 6 of *SMN1* and *SMN2*. *Human Genome Variation* 3: 16040.

Zeng, L, Pederson, SM, Kortschak, RD and Adelson, DL (2018). Transposable elements and gene expression during the evolution of amniotes. *Mobile DNA* 9: 17.

Figures

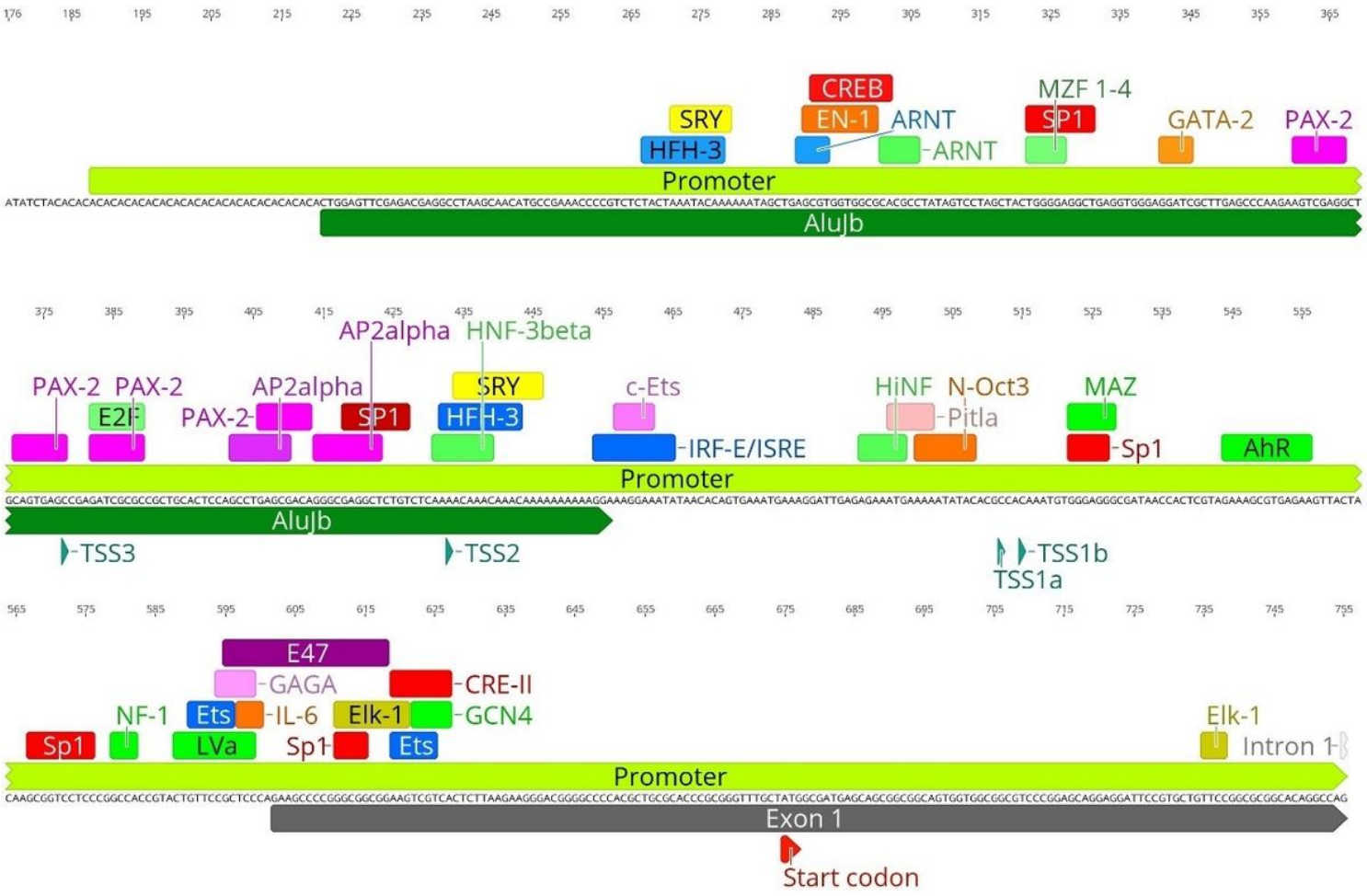

Figure 1

Identification of transcription factor binding sites in the AluJb element within the *SMN1/2* promoter region.

An AluJb element (represented by a dark green arrow) is inserted inside the promoter region of *SMN1/2*, upstream of the canonical transcriptional start site (TSS1a). Colored boxes indicate the locations of putative promoter elements and/or transcription factor binding sites, several located inside (as parts of) the AluJb sequence. Transcriptional start sites (TSS) and the start codon are represented by green and red arrowheads, respectively.

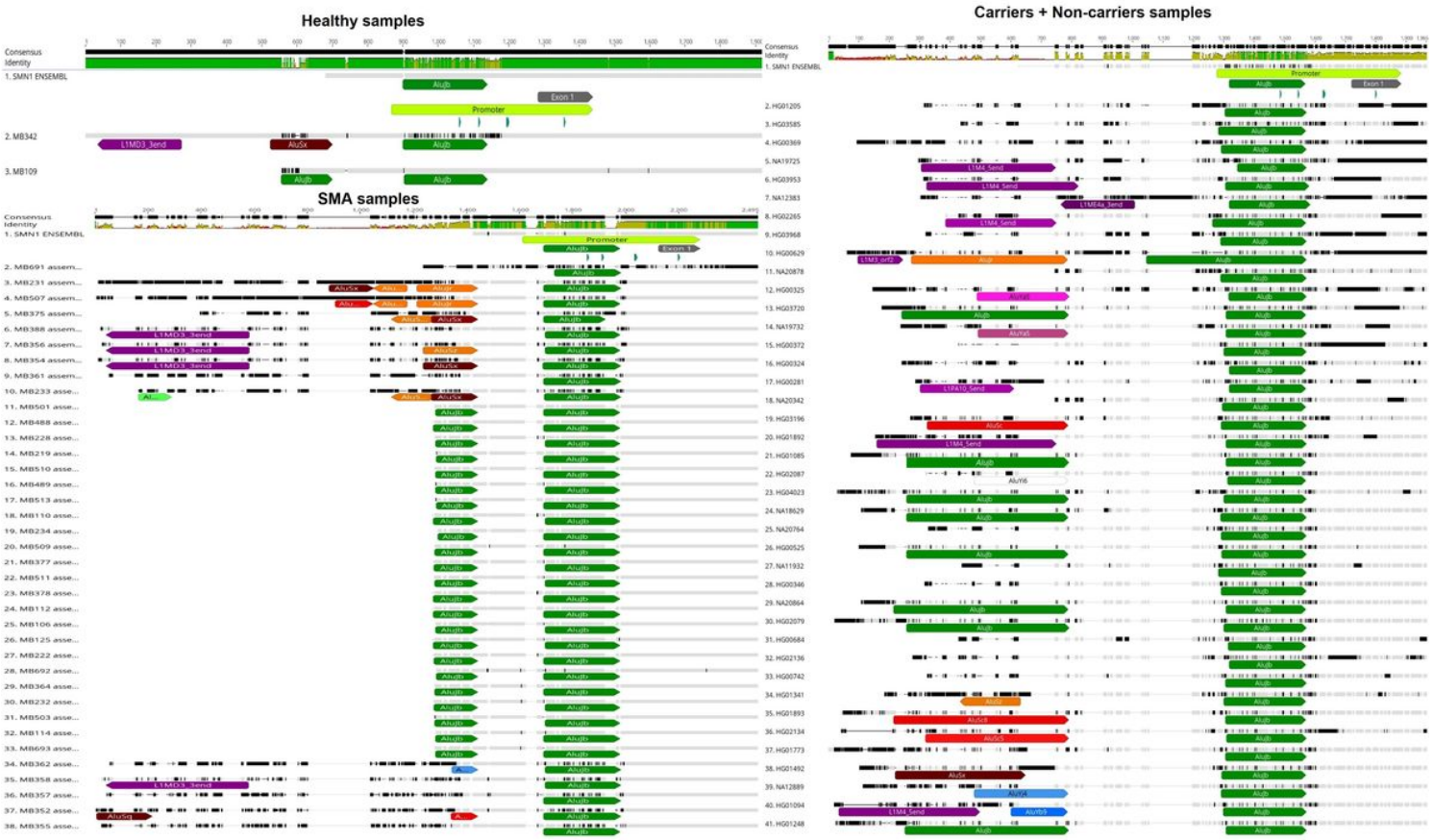

Figure 2

Comparison of the *SMN1*/*SMN2* promoter regions within sample cohort.

An AluJb insertion (represented by a dark green arrow) is present within the gene promotor region of all samples. Some polymorphic insertions were detected upstream of the promotor region and the 5'UTR; there were no connections, however, between these polymorphic insertions and SMA phenotype.

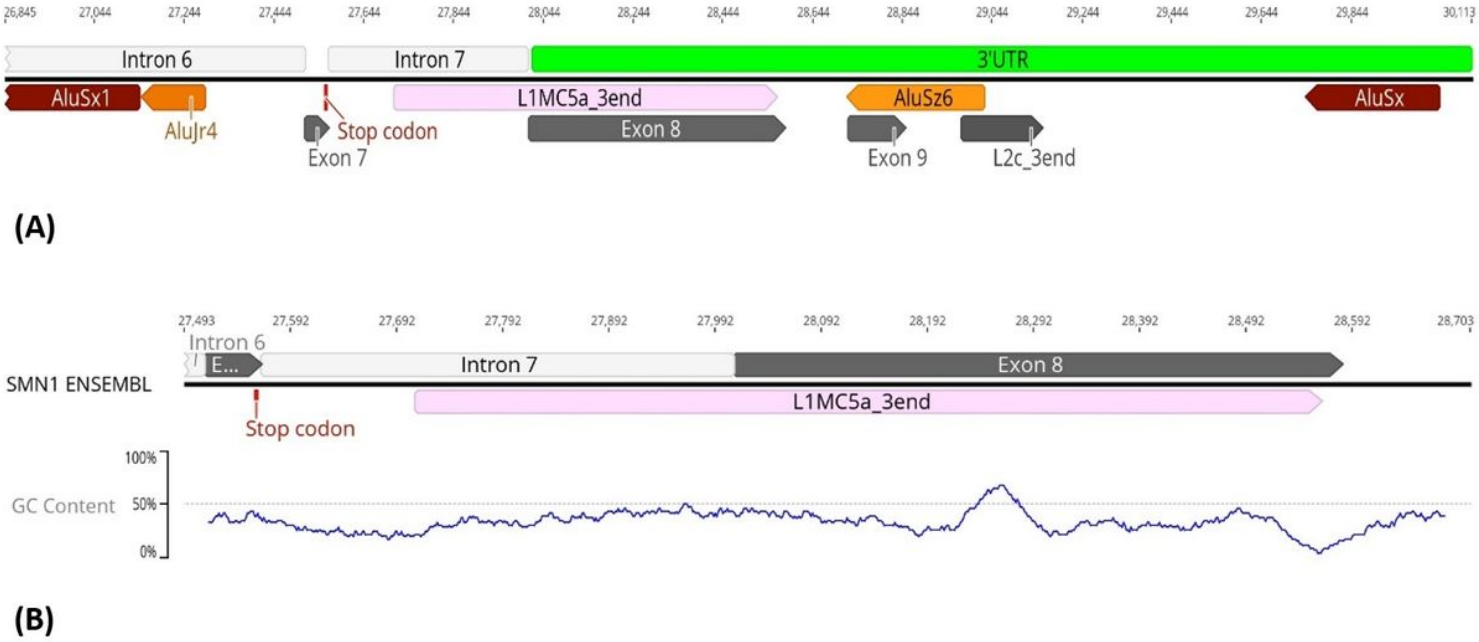

Figure 3

Identification of TEs within the *SMN1/2* 3'UTR.

(A) *SMN1/2* 3'UTR region is enriched in transposable elements (represented in colored arrows). Exon 8 is considered part of the 3'UTR region of the gene. An L1MC5a element (pink arrow) is inserted in *SMN1/2* exon 8 (grey arrow). **(B)** A large L1 insertion (represented by a colored arrow) was detected in *SMN1/2* last canonical exon, exon 8 (represented by a grey arrow). G+C analysis of the region showed a general lower G+C content in this region compared with the adjacent gene regions, partially explaining how a L1 insertion occurred in this region.

Figure 4

Comparison of *SMN1/2* exon 8 within the sample cohort.

All samples in study, independently of disease status, exhibit the L1 insertion (represented by a pink arrow) inside exon 8 (represented by a grey arrow) suggesting that the L1 element inserted in exon 8 is indeed fixed in the population and that it has a biological role in *SMN1/2* regulation.

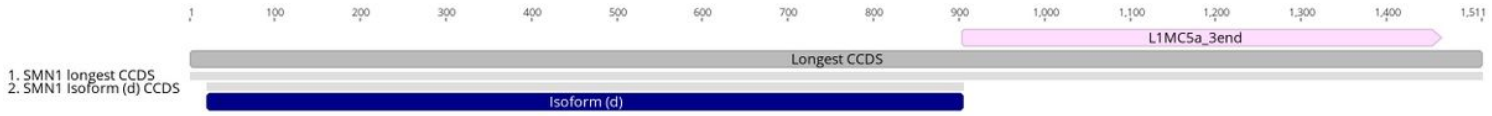

Figure 5

Identification of TEs within the *SMN1* Consensus Coding Sequences (CCDSs).

SMN1 longest isoform CCDS represented in grey on top has an L1MC5a element (represented by a pink arrow) inserted in its sequence responsible for the extension of the CCDS. Contrarily, *SMN1* most common CCDS, isoform (d; represented as a blue bar), is shorter and does not have any TE insertion in its sequence. Both sequences show 100% sequence identity within the overlapped region.

Figure 6

Identification of TEs within the reference *SMN1/2* gene sequence.

SMN1/2 exons are represented by grey arrows and introns by white boxes. *SMN1/2* promoter is represented by a green box and other regulatory motifs by green arrows. Start and stop codons are represented by small red arrows. Transposable elements position and orientation is indicated by colored arrows, with the direction of the arrow indicating the orientation of the repeat element.

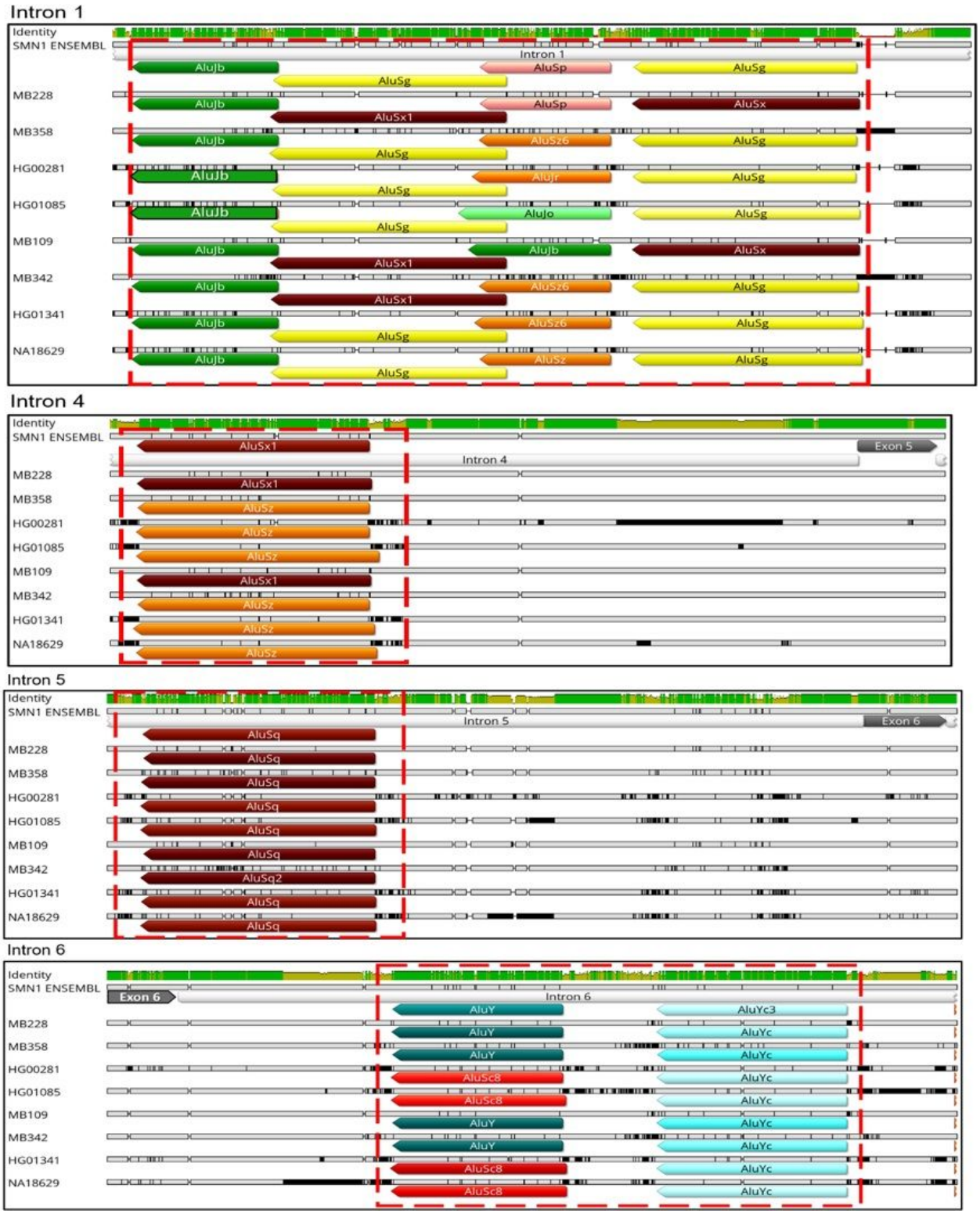

Figure 7

Identification of Key Alu repeats involved in RNA circularization events within *SMN1/2*.

Comparison between key Alu repeats involved in *SMN1/2* circularization events (inside the red boxes) located in introns 1, 4, 5 and 6. Independently of disease status, a conservation of position and orientation of the Alu insertions is visible. SMA patient samples (MB228 and MB358); SMA carriers

(HG00281 and HG01085); Healthy samples (MB109 and MB342) and Non-carriers samples (HG01341 and NA18629).

Figure 8

Identification of *SMN1/2* alternative exonization events.

On top, exonization event of an intronic antisense Alu repeat (represented by a green arrow) that gave birth to alternative exon 6B. Below, another exonization event of an antisense Alu element (represented by an orange arrow) that resulted in the formation of *SMN1/2* alternative exon 9. G+C content analysis of both regions shows a higher G+C content in the exonization regions when compared with the surrounding areas, which might have favored Alu insertions and the posterior exonization events.

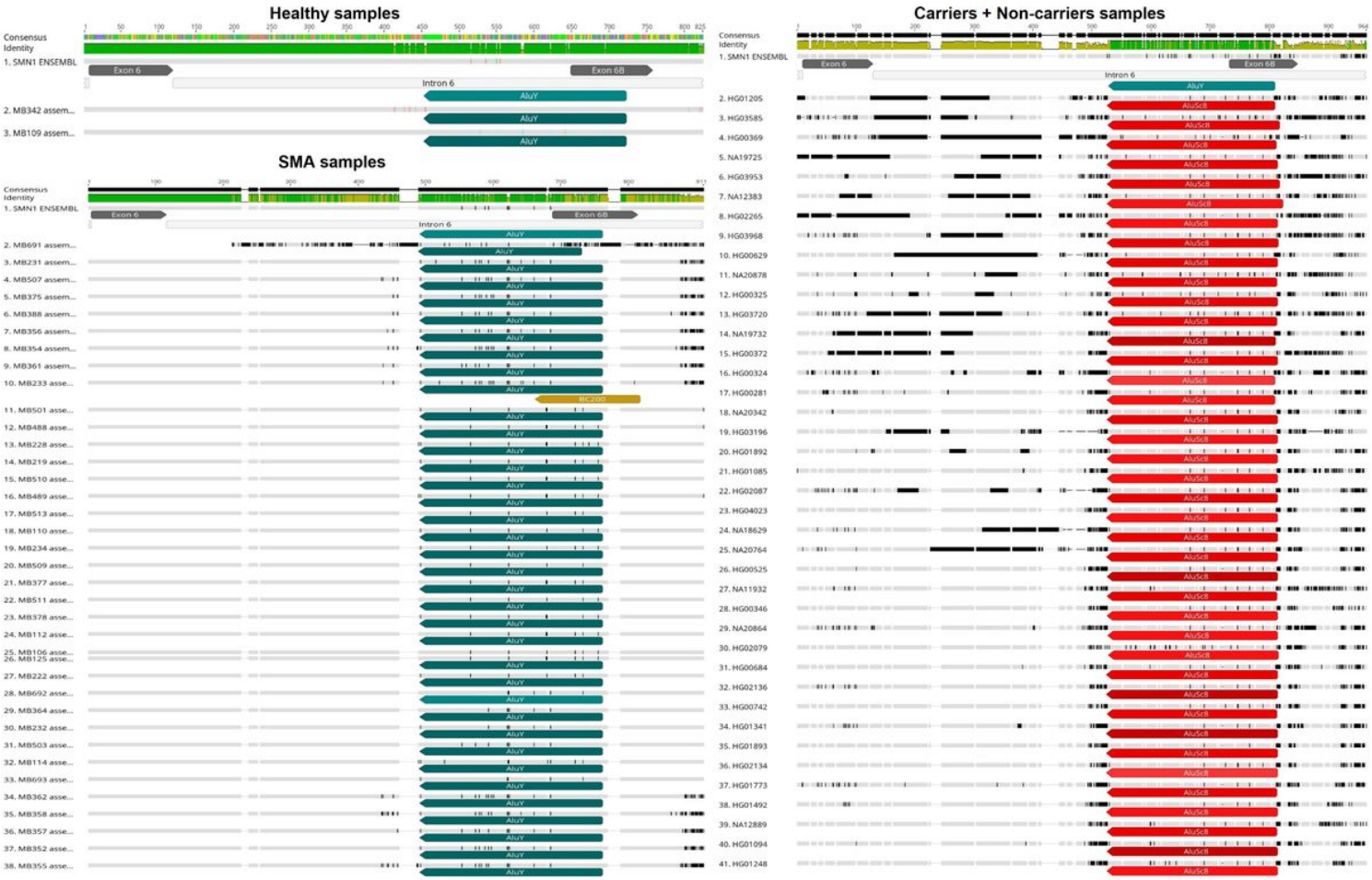

Figure 9

Comparison of the exon 6B region within the sample cohort.

An AluY insertion (represented by a green arrow) gives rise to alternative exon 6B in all healthy and SMA samples, including the reference *SMN1* ENSEMBL sequence. Contrarily, an AluSc8 (represented by a red arrow) insertion is present instead of the AluY element in the remaining samples. This Alu insertion difference is most likely the result of low sequencing read depth of the 1000 Genomes Project samples.

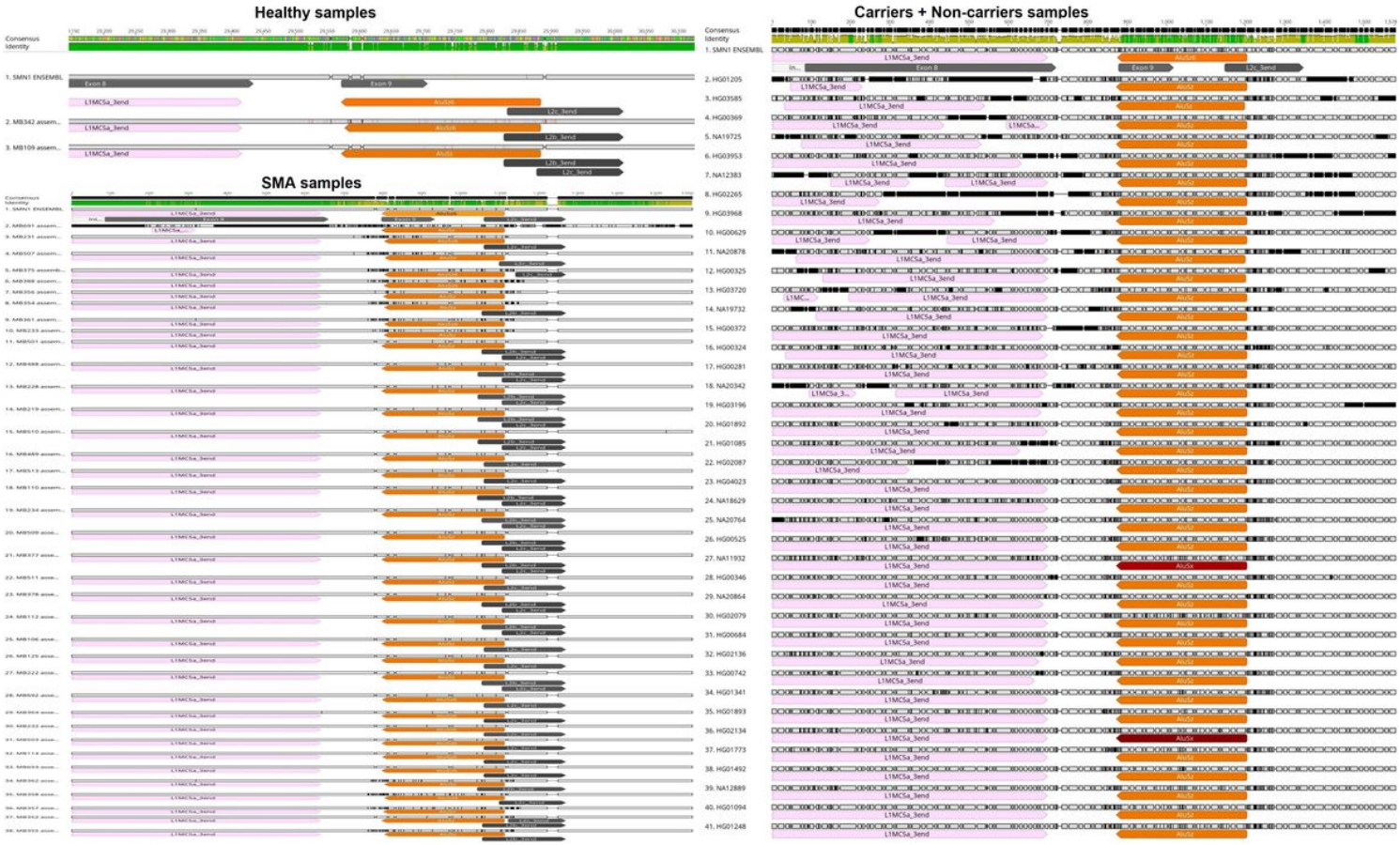

Figure 10

Comparison of the exon 9 region within the sample cohort.

Extensive insertional polymorphisms were detected for exon 9 region in the analyzed samples. While the expected AluSz insertion (represented by a light orange arrow) was present in some samples, other AluSz6 and AluSx insertions (represented by an orange arrow and a dark red arrow, respectively) were observed in this region. This polymorphism may be due to interpersonal variability and is not associated with SMA.

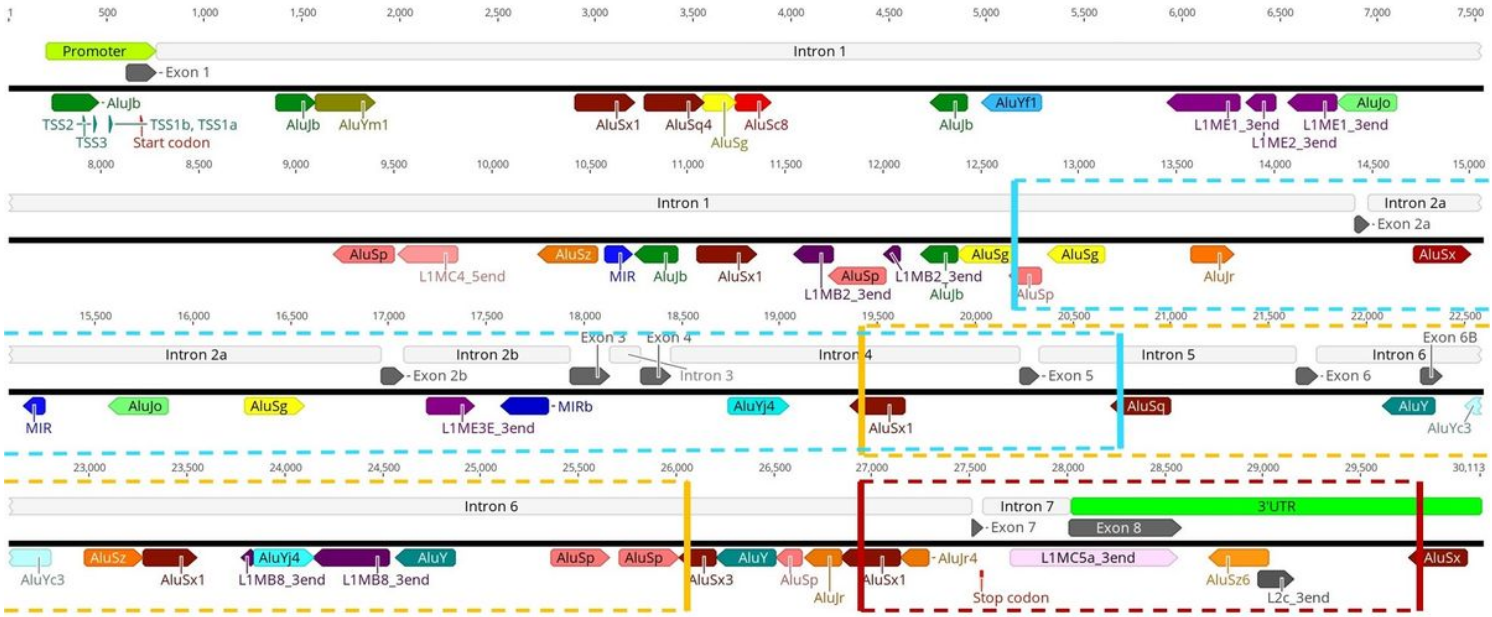

Figure 11

Identification of Alu-mediated partial deletions within *SMN1*.

The several Alu-derived repeats existent in *SMN1* introns provide a fertile source of Alu/Alu recombination events, known to lead to gene deletions. To date, three Alu-mediated deletion events were reported in *SMN1*. The more common deletion involving exon 7 and 8 is represented by a red box. A deletion event first described by Wirth *et al*. (50) involving exons 5 and 6 is indicated by a yellow box. Lastly, the more recently reported Alu-mediated deletion in *SMN1* is highlighted by a blue box.