SCISPACE
formerly Typeset

Open access • Posted Content • DOI:10.1101/2020.01.30.926790

RNA secondary structure mediated by Alu insertion as a novel disease-causing mechanism — Source link

Emmanuelle Masson, Sandrine Maestri, David Neil Cooper, Claude Férec ...+1 more authors

Institutions: French Institute of Health and Medical Research, Cardiff University

Published on: 31 Jan 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Alu element, Intron and RNA splicing

Related papers:

- Alteration of Splicing Pattern on Angiotensin Converting Enzyme Gene Due To The Insertion of Alu elements
- Multifactorial Interplay Controls the Splicing Profile of Alu-Derived Exons
- Exonization of AluYa5 in the human ACE gene requires mutations in both 3′ and 5′ splice sites and is facilitated by a conserved splicing enhancer
- The first reported case of Menkes disease caused by an Alu insertion mutation.
- Intronic Alus Influence Alternative Splicing

Share this paper:

View more about this paper here: https://typeset.io/papers/rna-secondary-structure-mediated-by-alu-insertion-as-a-novel-4s8clje9fb

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA secondary structure mediated by *Alu* insertion as a novel disease-causing mechanism

Emmanuelle Masson[1,2], Sandrine Maestri[1,2], David N. Cooper[3], Claude Férec[1,2] & Jian-Min Chen[1]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, F-29200 Brest, France.
[2]CHRU Brest, Service de Génétique, Brest, France.
[3]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

*Correspondence:
Jian-Min Chen, INSERM U1078, Faculté de Médecine, Bâtiment E – 2ème étage - Bureau E201b, 22 avenue Camille Desmoulins, F-29238 BREST Cedex 3, France.
Email: jian-min.chen@univ-brest.fr

ABSTRACT

We have recently reported a homozygous *Alu* insertion variant (termed *Alu*_Ins) within the 3'-untranslated region (3'-UTR) of the *SPINK1* gene as the cause of a new pediatric disease entity. Although *Alu*-Ins has been shown, by means of a full-length gene expression assay (FLGEA), to result in the complete loss of *SPINK1* mRNA expression, the precise underlying mechanism(s) has remained elusive. Herein, we filled this knowledge gap by adopting a hypothesis-driven approach. Employing RepeatMasker, we identified two *Alu* elements (termed *Alu*1 and *Alu*2) within the *SPINK1* locus; both are located deep within intron 3 and, most importantly, reside in the opposite orientation to *Alu*-Ins. Using FLGEA, we provide convincing evidence that *Alu*-Ins disrupts splicing by forming RNA secondary structures with *Alu*1 in the pre-mRNA sequence. Our findings reveal a previously undescribed disease-causing mechanism, resulting from an *Alu* insertion variant, which has implications for *Alu* detection and interpretation in human disease genes.

KEYWORDS

Alu insertion variant, human genetic disease, inverted *Alu* elements, L1 retrotransposition, pre-mRNA splicing, RNA secondary structure, template switching during reverse transcription

INTRODUCTION

Alu elements, about 300 nucleotides in length, are one of the most prevalent mobile elements in primate genomes. There are over one million copies in the human genome, accounting for almost 11% of the genome size[1]. *Alu* elements continue to be amplified in the human genome through long interspersed element-1 (LINE-1 or L1)-mediated retrotransposition[2,3], which is also an important cause of human genetic disease[3-5]. In part due to detection bias[6], disease-causing *Alu* insertions have invariably been found to be located within the coding or proximal intronic regions of affected genes until very recently, when a full-length *Alu* insertion (henceforth termed *SPINK1 Alu*-Ins) was identified in the 3'-untranslated region (3'-UTR) of the *SPINK1* gene (OMIM# 167790) in a patient presenting with a new pediatric disease entity, termed severe infantile isolated exocrine pancreatic insufficiency (SIIEPI)[7]. This new finding would not have come to light had the gene's 3'-UTR not been included in the mutational screen and had the *Alu*-Ins variant not been present in the homozygous state.

Pathological mechanisms underlying disease-causing *Alu* insertions (apart from those associated with concurrently generated large genomic deletions due to L1-mediated target-primed reverse transcription[8]) that have occurred within coding or proximal intronic regions have often not been explored experimentally but are generally thought either to involve the disruption of the coding sequences directly or to the promotion of aberrant splicing by virtue

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of their location[4,5]. We have previously analyzed the functional impact of the *SPINK1 Alu*-Ins event using a cell culture-based full-length gene expression assay (FLGEA). We established that it caused a complete loss of *SPINK1* mRNA expression in transfected HEK293T cells, an observation which was corroborated by three lines of evidence[7]. First, reverse transcription-PCR (RT-PCR) of mRNA from the *SPINK1 Alu*-Ins homozygote-derived cultured lymphocytes yielded no *SPINK1* transcripts. Second, a homozygous deletion of the entire *SPINK1* gene was found in a second SIIEPI patient. Third, significant pathological similarities in the pancreas were noted between SIIEPI patients and mice deficient for *Spink3*[9], the murine orthologue of *SPINK1*. This notwithstanding, the precise mechanism(s) underlying the complete loss of *SPINK1* mRNA expression due to *SPINK1 Alu*-Ins has remained to be elucidated. In the present study, we investigated the underlying mechanism by employing a hypothesis-driven approach, leading to the discovery of a previously undescribed disease-causing mechanism mediated by an *Alu* insertion variant.

RESULTS AND DISCUSSION

3'-UTR luciferase reporter assay

The 3'-UTRs of human genes play an important role in regulating mRNA 3' end formation, stability/degradation, nuclear export, subcellular localization and translation[10,11]. We surmised that the *SPINK1 Alu*-Ins mutation may affect one or more of these processes due to its location and therefore performed a 3'-UTR luciferase reporter assay. The *Alu*-Ins-containing 3'-UTR of the *SPINK1* gene was associated with a ~50% reduction in luciferase reporter activity as compared to its wild-type counterpart in both transfected HEK293T and COLO-357 cells (**Fig. 1**). This partial reduction failed to account for the previously observed complete functional loss of *SPINK1*[7], obliging us to look for other potential mechanisms.

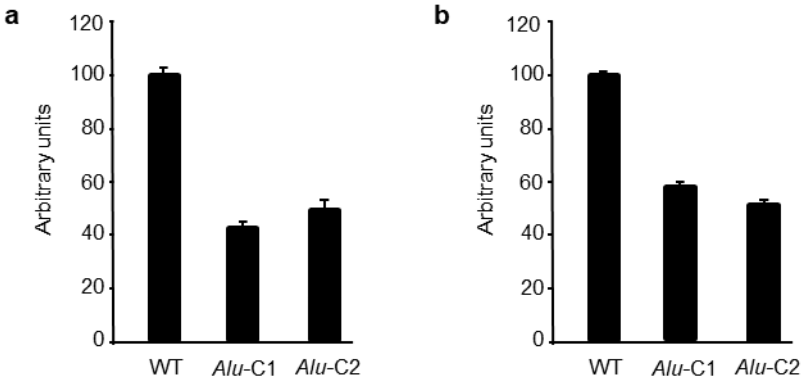

Figure 1. 3'-UTR luciferase reporter assay. Transfections were performed in both HEK293T (**a**) and COLO-357 (**b**) cells. WT, wild-type *SPINK1* 3'-UTR luciferase reporter vector. *Alu*-C1 and *Alu*-C2, two clones of the mutant *SPINK1* 3'-UTR luciferase reporter vector. Results were expressed as the mean ± S.D. of three independent experiments each performed in triplicate.

Hypothesis that the disease-causing *Alu*-Ins may form RNA secondary structure with pre-existing *Alu* element(s) within the *SPINK1* gene

Circular RNAs (circRNAs), an emerging class of RNA[12], are formed through a back-splicing mechanism[13]. Back-splicing is potentiated by secondary structures in the pre-mRNAs, and these secondary structures are often generated by inverted *Alu* elements[14-16]. Further, two artificial *Alu* elements inserted into an intron of a three-exon-minigene in opposite orientation have been shown experimentally to undergo base-paring thereby affecting the splicing pattern of the downstream exon[17]. These findings prompted us to search for the possible presence of *Alu* elements within the *SPINK1* locus that, if present, might have the potential to form RNA secondary structures with the disease-causing *Alu*-Ins. Employing RepeatMasker (http://www.repeatmasker.org/), we identified two such *Alu* elements within the *SPINK1* locus. Both are located deep within intron 3 of the *SPINK1* gene and, most importantly, both reside in the opposite orientation to *Alu*-Ins (**Fig. 2a**; **Supplementary Fig. 1)**. We surmised

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

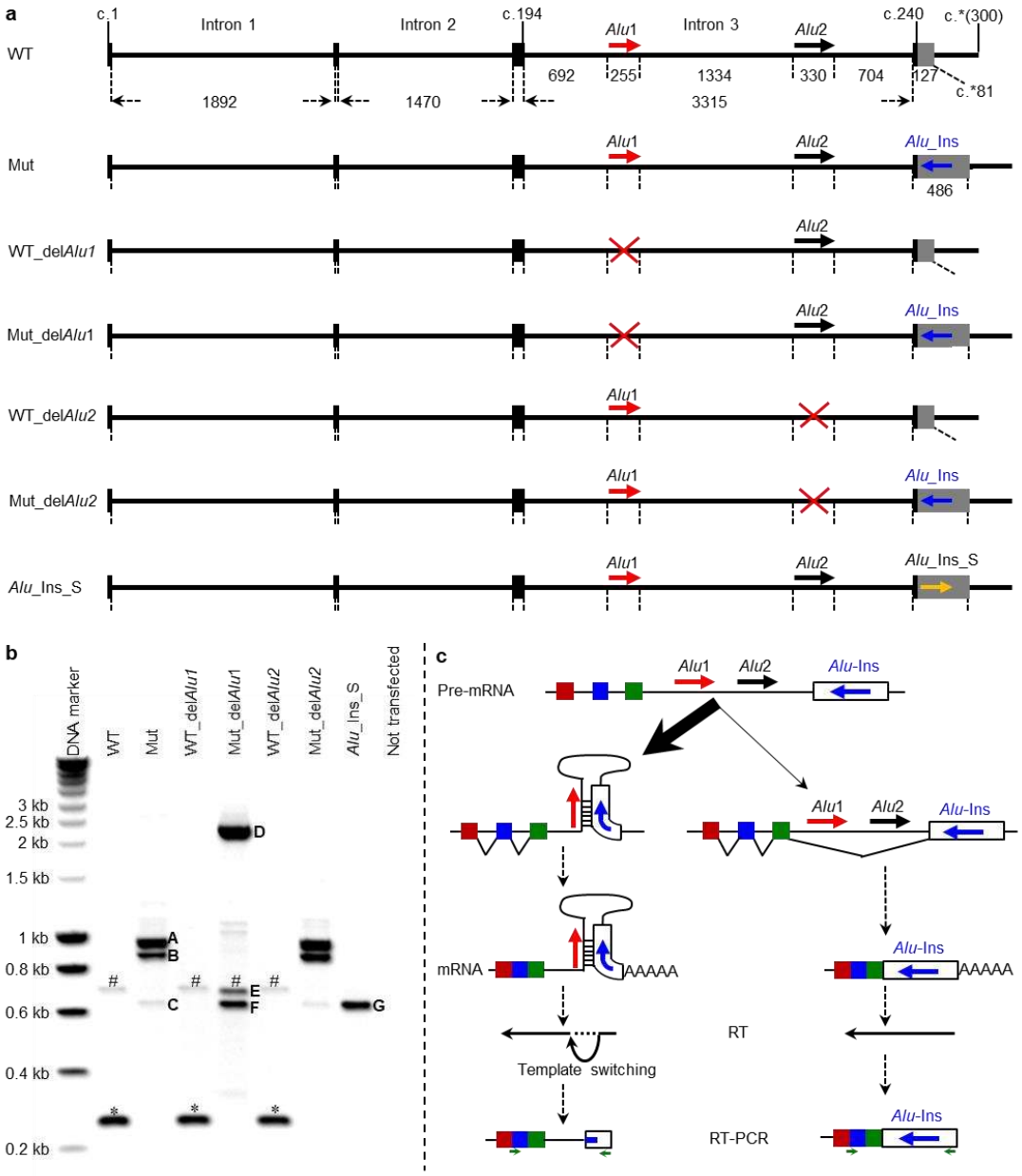

Figure 2. Full-length gene expression assay (FLGEA). (**a**) Illustration of the wild-type (WT), mutant (Mut) and five artificial *SPINK1* genomic sequences (drawn to scale) that were cloned into the pcDNA3.1/V5-His TOPO vector for the purposes of FLGEA. All inserts invariably spanned c.1 to c.*(300) by reference to the WT *SPINK1* genomic sequence. Coding sequences are shown as black bars and the 3'-UTR of *SPINK1* is shown as a grey box. Sizes (bp) of the different introns as well as different components of intron 3 are indicated. c.194, c.240 and c.*81 refer to the last nucleotide positions of exon 3, the translation termination codon and 3'-UTR, respectively. Orientations of the *Alu* elements are indicated by horizontal solid arrows. Red crosses indicate artificially deleted *Alu* sequences. *Alu*_Ins, the disease-causing *Alu* insertion. *Alu*_Ins_S, the artificially inverted version of *Alu*_Ins. (**b**) RT-PCR analyses of HEK293T cells transfected with the different expression vectors. The nature of the three bands in Mut_del*Alu*2 is similar to that of the three bands in Mut. Bands with * denote WT *SPINK1* transcripts. Bands with # denote non-specific amplifications. Bands C and F correspond to normally spliced transcripts containing *Alu*_Ins whereas band G corresponds to normally spliced transcripts containing *Alu*_Ins_S. Bands A, B and D correspond to template switching events (see Fig. 3) due to inverted *Alu*-mediated secondary structures. (**c**) A schema (not drawn to scale) to explain the generation of normally spliced transcripts (right panel) and template switching events (left panel) using the findings in Mut as an example. The facing green arrows indicate the approximate positions of the primers used for RT-PCR analysis; the forward primer spanned the exon 1/exon 2 junction whilst the reverse primer corresponded to nucleotides c.*29_*46 of the 3'-UTR of the *SPINK1* gene.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that *SPINK1 Alu*-Ins may have formed secondary structures with either or both of these two pre-existing *Alu* elements in the mutant pre-mRNA, thereby hindering the recognition of the 3' splice site of intron 3. However, the aberrantly spliced transcripts would have to have been degraded by mRNA decay mechanisms such as nonsense-mediated mRNA decay (NMD)[18], in order to explain the non-detection of *SPINK1* mRNA sequences from either mutant-transfected HEK293T cells or patient-derived lymphocytes observed in our previous study[7].

FLGEA re-performed in the presence of cycloheximide

To test the above hypothesis, we firstly re-performed the cell culture-based FLGEA for the previously constructed *SPINK1* wild-type (WT) and *Alu*-Ins mutant (Mut) expression vectors[7] but this time in the presence of cycloheximide, a known NMD inhibitor. RT-PCR of mRNAs from HEK293T cells transfected with the WT construct yielded two bands (**Fig. 2b**). The major band was confirmed to be the WT *SPINK1* transcript by directly sequencing the RT-PCR products. The faint band appeared to correspond to non-specific amplification products since sequencing the corresponding gel-purified product yielded no readable sequence despite multiple attempts. RT-PCR of mRNAs from cells transfected with the Mut construct yielded three bands (indicated as A, B and C in **Fig. 2b**). To identify the nature of these bands, we cloned the RT-PCR products into the pcDNA3.1/V5-His TOPO vector, transformed the resulting constructs into *E. coli*, and performed colony PCR followed by sequencing. The results of this experiment, including the number of colonies sequenced, nature of the corresponding RT-PCR products and their assignment to the three bands, are summarized in **Fig. 3a** (left panel).

a

Colony PCR analysis of RT-PCR products from HEK293T cells transfected with the Mut construct

Number of colonies sequenced (n=31)	Nature of sequence	Band assignment
2	A1	A
3	A2	A
2	A3	A
1	B1	B
6	B2	B
15	Correctly spliced	C
2	Not readable	—

Colony PCR analysis of RT-PCR products from HEK293T cells transfected with the Mut-del*Alu*1 construct

Number of colonies sequenced (n=30)	Nature of sequence	Band assignment
2	D1	D
6	D2	D
12	Correctly spliced	F
10	Not readable	—

b

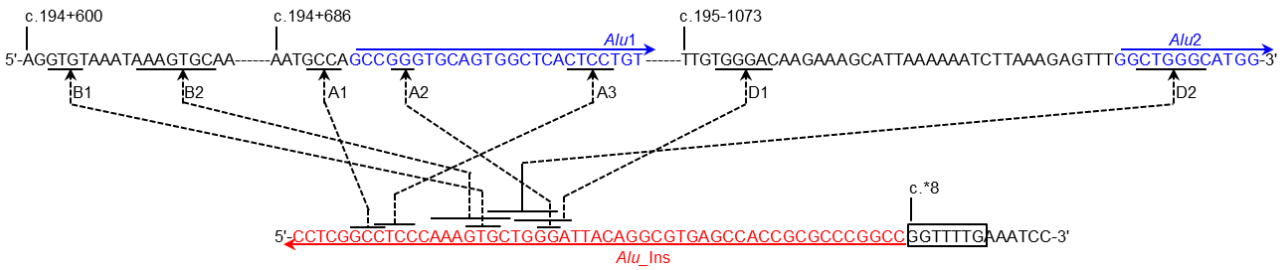

Figure 3. Findings from colony PCR followed by sequencing analysis. (**a**) Results obtained from analysis of the RT-PCR products derived from cells transfected with either the Mut construct (left panel) or Mut_del*Alu*1 construct (right panel). See Fig. 2b for the indicated bands. (**b**) Illustration of the different templated switching events and their associated direct short repeats. The boxed GGTTTG is the target site duplication.

Sequences obtained from more than half (n = 15) of the 29 Mut-derived informative colonies corresponded to correctly spliced transcripts and were assigned to band C (**Fig. 3a**). Since band C was only barely detectable, this high representation rate was probably attributable to a bias in ligation efficiency towards shorter inserts during the cloning of the RT-PCR products into the pcDNA3.1/V5-His TOPO vector. Moreover, the detection of barely detectable correctly spliced transcripts under the new experimental conditions suggested that cycloheximide may have an additional impact upon mRNA stability beyond its well-established role as an NMD inhibitor. Irrespective of the precise underlying mechanisms, our

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

new findings demonstrated that a minor fraction of mutant pre-mRNA sequences could undergo normal splicing (**Fig. 2c**).

Sequences obtained from the other 14 Mut-derived informative colonies corresponded to five distinct products, of which three (A1, A2 and A3) were assigned to band A whereas the other two (B1 and B2) were assigned to band B (**Fig. 3a**). All these products turned out to be template switching events, invariably involving a copy of the underlying short direct repeat within *SPINK1 Alu*-Ins and the other copy of the short direct repeat within the retained intron 3 sequence. Moreover, all copies of the involved short direct repeats within *SPINK1 Alu*-Ins are clustered within a 20-bp sequence tract whereas those within the retained intron 3 sequence are distributed over a <100-bp tract. Furthermore, whereas the former 20-bp sequence tract is very close to the 3' end of *SPINK1 Alu*-Ins, the latter <100-bp tract spans sequence immediately upstream of the 5' end of *Alu*1 (**Fig. 3b**). Based upon these sequence features, we reconstructed the following possible scenario: secondary structures were first formed between *Alu*1 and *SPINK1 Alu*_Ins in almost all mutant pre-mRNA sequences; these secondary structures prevented the splicing out of intron 3; template switching then occurred during reverse transcription of the secondary structure-containing mutant mRNAs[19]; finally, subsequent RT-PCR generated the band A and band B products (left panel in **Fig. 2c**).

FLGEA using five additional expression vectors

To provide further evidence to support the above postulate, we generated five additional expression vectors for FLGEA assay (**Fig. 2a; Supplementary Fig. 2**). Not surprisingly, deletion of either *Alu*1 or *Alu*2 in the context of the wild-type *SPINK1* sequence (WT_del*Alu*1 or WT_del*Alu*2) did not affect splicing; deletion of *Alu2* in the context of the mutant *SPINK1* sequence (Mut_del*Alu*2) yielded similar results as Mut (**Fig. 2b**). However, deletion of *Alu*1 in the context of the mutant *SPINK1* sequence (Mut_del*Alu*1) gave rise to three distinct RT-PCR bands (indicated as D, E and F in **Fig. 2b**). To identify the nature of these three bands, we performed colony PCR followed by sequencing as performed for the three Mut-derived bands. We found that band F comprised normally spliced transcripts whereas band D comprised two different products, D1 and D2 (right panel in **Fig. 3a**). As in the case of the faint band observed in WT, sequencing the gel-purified band E product yielded no readable sequence despite several attempts. D1 and D2 turned out to represent the products of template switching events, each involving a copy of the underlying short direct repeat within *SPINK1 Alu*-Ins and the other short direct repeat copy within the retained intron 3 sequence. Whereas copies of the D1- and D2-involved short direct repeats within *SPINK1 Alu*-Ins are located within the above-mentioned 20-bp sequence tract, those within the retained intron 3 sequence span sequence immediately upstream of the 5' end of *Alu*2. These sequence features suggested that in the absence of *Alu*1, *SPINK1 Alu*-Ins could also form secondary structures with *Alu*2, which in turn hinders the splicing out of intron 3. Nonetheless, given the detection of more abundant normally spliced transcripts from Mut-del*Alu*1 than from Mut (**Fig. 2b**), secondary structures formed between *Alu*2 and *SPINK1 Alu*-Ins would appear to be less stable than those formed between *Alu*1 and *SPINK1 Alu*-Ins. Finally, inversion of the *SPINK1 Alu*_Ins within the 3'-UTR of the *SPINK1* gene resulted in an artificial *Alu* insertion in the same orientation as *Alu*1 and *Alu*2 (*Alu*_Ins_S); RT-PCR analysis of HEK293T cells transfected with the corresponding expression vector yielded a single band (i.e., band G in **Fig. 2b)**, which was confirmed by sequencing to correspond to normally spliced transcripts. This finding clearly showed that it is not the primary sequence of the *Alu* insertion but rather its potentiated secondary structure that is critical for the virtually complete functional loss of the affected *SPINK1* allele.

CONCLUSIONS

In summary, employing FLGEA, we have provided convincing evidence that the *Alu* insertion into the 3'-UTR of the *SPINK1* gene disrupted splicing by forming secondary structures with a pre-existing *Alu* element that is located deep within intron 3 of the gene. Although inverted *Alu* elements have long been known to be capable of forming secondary structures, this is the first time that an *Alu* insertion variant has been shown to exert its effect via such a

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mechanism causing human genetic disease. Given the abundance of pre-existing *Alu* elements within the human genome and the potential insertion of new *Alu* elements into virtually any site, our findings have important implications for the detection and interpretation of *Alu* elements in human disease genes: it follows that not only coding and proximal intronic sequences but also deep intronic and non-coding regions should be analyzed for possible *Alu* insertions. The functional effect of *Alu* insertions should be considered with respect to their distances and orientations in relation to *Alu* elements that are present within the target genes.

METHODS

Reference sequence

SPINK1 mRNA reference sequence accession NM_003122.3 defines a five-exon gene, of which the first exon is non-coding. However, the gene expressed in the pathophysiologically relevant exocrine pancreas comprises only four exons[20,21], which correspond to NM_003122.3-defined exons 2-5. To date, this latter four-exon gene has been invariably used by the genetics field as the *de facto SPINK1* reference gene (see Tang and colleagues[22] and references therein). This convention was followed here, with the corresponding *SPINK1* reference genomic sequence being obtained from the GRCh38/hg38 assembly (https://genome.ucsc.edu/). Nucleotide numbering was based upon the coding DNA sequence according to Human Genome Variation Society (HGVS) recommendations[23].

3'-UTR luciferase reporter assay

Construction of wild-type and mutant reporter vectors

The 3'-UTR of the wild-type (WT) *SPINK1* gene is 81-bp long (**Supplementary Fig. 1**). A 306-bp fragment containing the entire 3'-UTR (i.e., going from c.*1_*81) plus the downstream 225-bp 3' flanking sequence (i.e., sequence from c.*(82)_*(306)) of the WT *SPINK1* gene was PCR amplified from genomic DNA of a healthy subject using forward primer 5'-TCTAGAGAACCAAGGTTTTGAAATCCCA-3' containing an *Xba*I restriction site (underlined) and reverse primer 5'-GGATCCGATCATCTGTGCTCTGCCAT-3' containing a *BamH*I restriction site (underlined). The corresponding fragment containing the disease-causing *Alu* insertion, *SPINK1 Alu*_Ins, was amplified from genomic DNA of the homozygous patient. The resulting PCR products were firstly digested with *Xba*I and *BamH*I and then cloned into the *Xba*I/*BamH*I sites of the pGL3 Control Vector (Promega, Charbonnieres, France), respectively. In the resulting pGL3-WT or pGL3-*Alu SPINK1* 3'-UTR reporter gene construct, the insert was placed immediately downstream of the translational termination codon of the luciferase reporter gene. The accuracy of both inserts was verified by sequencing using the BigDye™ Terminator v1.1 Cycle Sequencing Kit (ThermoFisher Scientific, Illkirch, France).

Cell culture, transfection and luciferase reporter gene assay

Human embryonic kidney (HEK293T) and human pancreatic adenocarcinoma (COLO-357) cell lines were maintained in DMEM nutrient mixture supplemented with 10% fetal calf serum. Transfections were carried out as previously described[24] using 3.8 µg pGL3-WT or pGL3-*Alu SPINK1* 3'-UTR luciferase reporter vector plus 0.2 µg control pRL-CMV vector. At 48 h after transfection, luciferase measurement was conducted as previously described[25].

Full-length gene expression assay (FLGEA)

Construction of an additional five expression vectors

Full-length WT and mutant (Mut) *SPINK1* expression vectors were previously constructed[7]. Five additional expression vectors for FLGEA assay were newly generated in the present study.

To generate an *Alu1* deletion in the context of the WT or mutant *SPINK1* sequence, two fragments (indicated by the primer pairs in different colors; **Supplementary Fig. 2a**) were amplified using genomic DNA from a healthy control or the *SPINK1 Alu*_Ins homozygous patient. For each fragment, PCR reaction was performed with 50 ng DNA in a 50 µL reaction mixture containing 2.5 U TaKaRa La Taq™ DNA polymerase, 400 µM TakaRa dNTP Mix,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and 0.4 µM each of the corresponding primer pair. The PCR program comprised an initial denaturation at 94°C for 1 min, 35 cycles of denaturation at 94°C for 20 s, annealing at 56°C for 20 s and extension at 72°C for 5 min, and a final extension step at 72°C for 10 min. After digestion with *Kpn*I, the two fragments obtained from either subject were ligated together. Each ligated product was then amplified using forward primer P1_F and reverse primer P1_R (**Supplementary Fig. 2a**). The PCR was performed using the GoTaq® Long PCR MasterMix (Promega, Charbonnieres, France) according to the manufacturer's protocol. The PCR program had an initial denaturation at 94°C for 2 min, 35 cycles of denaturation at 94°C for 20 s, annealing at 58°C for 20 s and extension at 65°C for 6 min, and a final extension step at 72°C for 10 min. The resulting two PCR products, WT_del*Alu*1 and Mut_del*Alu*1, were then separately cloned into the pcDNA3.1/V5-His TOPO vector according to the manufacturer's instructions.

Expression vectors for *Alu*2 deletions in both the WT and mutant *SPINK1* sequence contexts, WT-del*Alu*2 and Mut_del*Alu*2, were constructed essentially as for WT_del*Alu*1 and Mut_del*Alu*1. The only differences were the use of different primers (**Supplementary Fig. 2b**) and different extension times for amplifying the two fragments (6 min for the longer one (amplified by P1_F and *Alu*2_*Kpn*I_R) and 1 min for the shorter one (amplified by *Alu*2_*Kpn*I_F and P1_R)).

To construct an expression vector for *Alu*_Ins_S (i.e., inversion of the disease-causing *Alu*_Ins within 3'-UTR of the *SPINK1* gene), a natural *Nco*I restriction enzyme site located within intron 3 of the *SPINK1* gene in the context of the previously constructed Mut expression vector was firstly eliminated by means of the QuickChange II XL Site-Directed Mutagenesis Kit (Agilent, Les Ulis, France) according to the manufacturer's instructions; the primers used were 5'-TGGCCAACATGGTGAAACCCCGTGGTGGCGGGCGCCTATAATAC-3' (forward) and 5'-GTATTATAGGCGCCCGCCACCACGGGGTTTCACCATGTTGGCCA-3' (reverse). Then three PCR reactions were performed each with 1 ng of the modified plasmid in a 50 µL reaction mixture containing 2.5U TaKaRa La Taq™ DNA polymerase, 400 µM TakaRa dNTP Mix, and 0.4 µM each of the corresponding primer pair (**Supplementary Fig. 2c**). The PCR program had an initial denaturation at 94°C for 1 min, 35 cycles of denaturation at 94°C for 20 s, annealing at 56°C for 20 s and extension at 72°C for 7 min (fragment A) or 1 min (both fragments B and C), and a final extension step at 72°C for 10 min. After being digested by *Kpn*I and/or *Nco*I, the three fragments were ligated together. The subsequent steps were the same as described above for the other constructs.

Primer sequences are provided in **Supplementary Fig. 2d**. All exon/intron boundaries and ligation junction(s) of the newly generated inserts were verified by sequencing.

Cell culture, transfection and reverse transcription

HEK293T cells were maintained as described earlier. Transfections were carried out using 1 µg of the expression plasmid per well in a 6-well plate. Four hours before the RNA extraction, cells were treated with cycloheximide with a final concentration of 50 µg/ml as previously described[26]. At 24h after transfection, the cells were harvested for total RNA extraction using TRIzol RNA Isolation Reagents (ThermoFisher Scientific). The RNA concentration and purity were determined by measuring the OD at 260 nm and 280 nm, respectively. 4 µg RNA was treated with DNAse I (ThermoFisher Scientific, Illkirch, France) before reverse transcription (RT). RT was performed in a 20 µL mixture containing 1 µg RNA treated by DNase I, 10 U RNAse inhibitor (Promega, Charbonnieres, France), 250 ng Random Hexamers (Qiagen, Courtaboeuf, France), 4 µl 5× First Strand Buffer, 500 µM dNTPs, 5mM dithiothreitol and 200 U SuperScript® II Reverse Transcriptase (ThermoFisher Scientific, Illkirch, France). The reaction was incubated at 42°C for 50 min and inactivated by heating at 70°C for 15 min. The resulting complementary DNA (cDNA) were treated with 2U RNaseH (ThermoFisher Scientific, Illkirch, France) to degrade the remaining RNA at 37°C for 20 min.

RT-PCR

RT-PCR was performed using forward primer 5'-GAGTCTATCTGGTAACACTGGAGCT-3' and reverse primer 5'-CAGTCAGGCCTCGCGGTG-3' and the GoTaq® Long PCR

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MasterMix according to the manufacturer's protocol. The PCR program had an initial denaturation at 94°C for 2 min, followed by 40 cycles of denaturation at 94°C for 20 s, annealing at 58°C for 20 s, extension at 65°C for 4 min, and a final extension step at 72°C for 5 min. PCR products were evaluated by electrophoresis on an 1% agarose gel.

Sequencing of the RT-PCR products

RT-PCR products resulting from the WT, WT_del*Alu*1, WT_del*Alu*2 and *Alu*_Ins_S expression vectors were respectively purified by Illustra™ ExoProStar™ (GE Healthcare, Orsay, France) and directly sequenced by means of the BigDye™ Terminator v1.1 Cycle Sequencing Kit using forward primer 5'-GAGTCTATCTGGTAACACTGGAGCT-3' and reverse primer 5'-CAGTCAGGCCTCGCGGTG-3'.

RT-PCR products resulting from the Mut, Mut_del*Alu1* and Mut_del*Alu2* expression vectors were separately cloned in the pcDNA3.1/V5-His TOPO vector according to the manufacturer's instructions. Transformation was performed using XL10-Gold Ultracompetent Cells (Stratagene, La Jolla, CA). Transformed cells were spread onto LB agar plates with 50 mg/mL ampicillin and incubated at 37ºC overnight. Some 30 colonies from each plate were subjected to PCR amplification using the GoTaq® Long PCR MasterMix according to the manufacturer's protocol; the primers used were 5'-GGAGACCCAAGCTGGCTAGT-3' (forward) and 5'-AGACCGAGGAGAGGGTTAGG-3' (reverse), both being located within the vector sequence. The PCR program had an initial denaturation at 94°C for 2 min, followed by 35 cycles of denaturation at 94°C for 20 s, annealing at 58°C for 20 s, extension at 65°C for 1 min (Mut and Mut_del*Alu2*) or 4 min (Mut_del*Alu1*), and a final extension step at 72°C for 5 min. PCR products were controlled by electrophoresis on an 1% agarose gel, purified by Illustra™ ExoProStar™ and sequenced using the BigDye™ Terminator v1.1 Cycle Sequencing Kit. The primers used for sequencing were 5'-GGAGACCCAAGCTGGCTAGT-3' (forward), 5'-TGAAAATCGGTGAGTACA-3' (forward), 5'-GAAAACATCATGAGCATG-3' (forward) and 5'- AGACCGAGGAGAGGGTTAGG-3' (reverse). For Mut_del*Alu1*-derived products, the larger band (>2 kb) was further sequenced using three additional sequencing primers (all forward): 5'-CTGAGATTGACTTGAT-3', 5'-TCTGAAACCTCCGAGT-3' and 5'-CTAACTTAAATGTGGCT-3'.

Bands whose nature remained undermined after the aforementioned sequencing efforts were excised from the agarose gel, purified by MinElute Gel extraction kit (Qiagen, Courtaboeuf, France) and sequenced with the BigDye™ Terminator v1.1 Cycle Sequencing Kit. The primers used were 5'-GAGTCTATCTGGTAACACTGGAGCT-3' (forward) and 5'-CAGTCAGGCCTCGCGGTG-3' (reverse).

DATA AVAILABILITY

All data generated or analyzed during this study are included in this manuscript and its supplementary information files.

ACKNOWLEDGMENTS

This study was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), France.

AUTHOR CONTRIBUTIONS

E.M. designed the study, performed the experiments, analyzed the data and contributed to figure preparation and paper writing. S.M. assisted in performing the experiments. D.N.C. critically revised the manuscript. C.F. provided funding and supervised the study. J.M.C. conceived and designed the study, analyzed the data, prepared figures and wrote the manuscript. All authors read and proved the final manuscript.

COMPETING INTEREST STATEMENT

The authors are unaware of any conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.926790; this version posted January 31, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**, 1567-1577 (2019).
3. Kazazian, H.H., Jr. & Moran, J.V. Mobile DNA in health and disease. *N Engl J Med* **377**, 361-370 (2017).
4. Chen, J.M., Stenson, P.D., Cooper, D.N. & Férec, C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117**, 411-27 (2005).
5. Hancks, D.C. & Kazazian, H.H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9 (2016).
6. Chen, J.M., Férec, C. & Cooper, D.N. LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. *J Biomed Biotechnol* **2006**, 56182 (2006).
7. Venet, T. *et al.* Severe infantile isolated exocrine pancreatic insufficiency caused by the complete functional loss of the *SPINK1* gene. *Hum Mutat* **38**, 1660-1665 (2017).
8. Callinan, P.A. *et al. Alu* retrotransposition-mediated deletion. *J Mol Biol* **348**, 791-800 (2005).
9. Ohmuraya, M. *et al.* Autophagic cell death of pancreatic acinar cells in serine protease inhibitor Kazal type 3-deficient mice. *Gastroenterology* **129**, 696-705 (2005).
10. Chen, J.M., Férec, C. & Cooper, D.N. A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. *Hum Genet* **120**, 1-21 (2006).
11. Steri, M., Idda, M.L., Whalen, M.B. & Orru, V. Genetic variants in mRNA untranslated regions. *Wiley Interdiscip Rev RNA* **9**, e1474 (2018).
12. Kristensen, L.S. *et al.* The biogenesis, biology and characterization of circular RNAs. *Nat Rev Genet* **20**, 675-691 (2019).
13. Welden, J.R. & Stamm, S. Pre-mRNA structures forming circular RNAs. *Biochim Biophys Acta Gene Regul Mech*, 194410 (2019).
14. Jeck, W.R. *et al.* Circular RNAs are abundant, conserved, and associated with *ALU* repeats. *RNA* **19**, 141-57 (2013).
15. Liang, D. & Wilusz, J.E. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev* **28**, 2233-47 (2014).
16. Zhang, X.O. *et al.* Complementary sequence-mediated exon circularization. *Cell* **159**, 134-147 (2014).
17. Lev-Maor, G. *et al.* Intronic *Alu*s influence alternative splicing. *PLoS Genet* **4**, e1000204 (2008).
18. Lindeboom, R.G.H., Vermeulen, M., Lehner, B. & Supek, F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat Genet* **51**, 1645-1651 (2019).
19. Cocquet, J., Chong, A., Zhang, G. & Veitia, R.A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**, 127-31 (2006).
20. Yamamoto, T. *et al.* Molecular cloning and nucleotide sequence of human pancreatic secretory trypsin inhibitor (*PSTI*) cDNA. *Biochem Biophys Res Commun* **132**, 605-12 (1985).
21. Horii, A. *et al.* Primary structure of human pancreatic secretory trypsin inhibitor (*PSTI*) gene. *Biochem Biophys Res Commun* **149**, 635-41 (1987).
22. Tang, X.Y. *et al.* Toward a clinical diagnostic pipeline for *SPINK1* intronic variants. *Hum Genomics* **13**, 8 (2019).
23. den Dunnen, J.T. *et al.* HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat* **37**, 564-9 (2016).
24. Boulling, A., Le Gac, G., Dujardin, G., Chen, J.M. & Férec, C. The c.1275A>G putative chronic pancreatitis-associated synonymous polymorphism in the glycoprotein 2 (*GP2*) gene decreases exon 9 inclusion. *Mol Genet Metab* **99**, 319-24 (2010).
25. Boulling, A. *et al.* Assessing the pathological relevance of *SPINK1* promoter variants. *Eur J Hum Genet* **19**, 1066-73 (2011).
26. Zou, W.B. *et al.* No association between *CEL-HYB* hybrid allele and chronic pancreatitis in Asian populations. *Gastroenterology* **150**, 1558-1560 e5 (2016).