A birth of bipartite exon by intragenic deletion

Kandai Nozu[1], Kazumoto Iijima[1], Toru Igarashi[2], Shiro Yamada[3,4], Jana Kralovicova[5], Yoshimi Nozu[1], Tomohiko Yamamura[1], Shogo Minamikawa[1], Ichiro Morioka[1], Takeshi Ninchoji[1], Hiroshi Kaito[1], Koichi Nakanishi[6], Igor Vorechovsky[5]

[1]Department of Pediatrics, Kobe University Graduate School of Medicine, Kobe, Japan
[2]Department of Pediatrics, Nippon Medical School Hospital, Tokyo, Japan
[3]Department of Pediatrics, Tokai University Oiso Hospital, Oiso, Kanagawa, Japan
[4]Division of Human Genetics, National Institute of Genetics, Mishima, Japan
[5]University of Southampton Faculty of Medicine, Southampton, United Kingdom
[6]Department of Pediatrics, Wakayama Medical University, Wakayama, Japan

Running title: Composite LINE-1 exonization in *COL4A5* in Alport syndrome

Corresponding authors:
Kandai Nozu, M.D., Ph.D.
Department of Pediatrics,
Kobe University Graduate School of Medicine,
7-5-1 Kusunoki-cho, Chuo, Kobe, Hyogo 6500017, Japan.
Tel.: +81-78-382-6090; Fax: +81-78-382-6099
Email: nozu@med.kobe-u.ac.jp

Igor Vorechovsky, M.D., Ph.D.
Faculty of Medicine
University of Southampton
HDH, MP808, Tremona Road
Southampton SO16 6YD, United Kingdom
Tel.: +44 2381 206425; Fax: +44 2381 204264
Email: igvo@soton.ac.uk

Abstract

Background

Disease-causing mutations that activate transposon-derived exons without creating a new splice-site consensus have not been previously reported, but they may provide unique insights into our understanding of structural motifs required for inclusion of intronic sequences in mature transcripts.

Methods

We employ a combination of experimental and computational techniques to characterize the first *de novo* bipartite exon activation in genetic disease.

Results

The bipartite exon resulted from an in-frame *COL4A5* deletion associated with a typical Alport syndrome. The deletion encompassed exons 38 through 41 and activated a cryptic 3’ and 5’ splice site that were derived from intron 37 and intron 41, respectively. The deletion breakpoint was in the middle of the new exon, with considerable reverse complementarity between the two exonic parts, potentially bringing the cryptic 3’ and 5’ splice site into proximity. The 3’ splice site, polypyrimidine tract and the branch site of the new exon were derived from an inactive, 5’ truncated LINE-1 retrotransposon. This ancient LINE-1 copy sustained a series of mutations that created the highly conserved AG dinucleotide at the 3’ splice site early in primate development. The exon was fully included in mature transcripts and introduced a stop codon in the shortened *COL4A5* mRNA, illustrating pitfalls of inferring disease severity from DNA mutation alone.

Conclusion

These results expand the repertoire of mutational mechanisms that alter RNA processing in genetic disease and illustrate the extraordinary versatility of transposed elements in shaping the new exon-intron structure and the phenotypic variability.

Key words: *COL4A5*, exon, transposon, deletion, LINE-1, Alport syndrome, splicing, branch site, RNA

Introduction

Hereditary diseases are often caused by intronic mutations that create new 3’ or 5’ splice sites to activate cryptic exons (Busslinger et al. 1981; Buratti et al. 2011). Mutation-induced activation of cryptic splice sites is common in transposed elements, such as *Alu*s or mammalian-wide interspersed repeats, contributing significantly to human morbidity, phenotypic variability and the evolution of exon-intron structure (Vorechovsky 2010; Schmitz and Brosius 2011). However, aberrant transcripts may also arise from intronic variants that alter auxiliary splicing sequences, known as enhancers and silencers (King et al. 2002; Pagani et al. 2002). Intronic splicing enhancers or silencers promote or inhibit recognition of cryptic splice sites, which have similar sequences as authentic splice sites but outnumber them by at least an order of magnitude in the genome (Fairbrother et al. 2000). Reports of disease-causing mutations that activate new exons without creating a new splice-site consensus have been extremely rare, yet they may provide unique insights into our understanding of sequence and structural motifs required for inclusion of introns in mature transcripts (Pagani et al. 2002; Buratti et al. 2007b; Vorechovsky 2010). However, disease-causing mutations that activate transposon-derived exons without creating a new splice-site consensus have not been previously reported.

Alport syndrome is characterized by a progressive kidney disease accompanied by hearing loss and ocular abnormalities (Kashtan 1999). About 85% patients with Alport syndrome are due to mutations in the *COL4A5* gene (X-linked Alport syndrome, XLAS, MIM:301050), which encodes the α5 chain of type IV collagen (Kashtan 1999). Over 900 different pathogenic *COL4A5* variants have been identified in XLAS, including large deletions, splice-site mutations and cryptic exon activation (King et al. 2002; Nozu et al. 2014a; Nozu et al. 2014b; Oka et al. 2014), but the XLAS mutation pattern is far from complete.

Here we describe a heterozygous *COL4A5* deletion that activated a bipartite cryptic exon, with its 5’ and 3’ splice sites derived from distinct introns. The 3’ splice site, polypyrimidine tract and a branch site were contributed by an inactive copy of the Long INterspersed Element (LINE-1), providing a new paradigm for the retrotransposon-mediated phenotypic variability.

Material and methods

Genomic DNA was isolated from peripheral blood leukocytes of the proband and both parents using the Quick Gene Mini 80 System (Fujifilm Corporation, Tokyo, Japan) according to the manufacturer's instructions. Targeted next-generation sequencing was carried out using the HaloPlex target enrichment system for the *COL4A3*, *COL4A4* and *COL4A5* genes (Agilent Technologies), employing MiSeq (Illumina) and SureCall (v. 3.0; Agilent Technologies). Multiplex Ligation-dependent Probe Amplification (MLPA) was performed using the SALSA P191/P192 Alport assay (V.04; MRC-Holland) according to supplier's recommendations. Total RNA was extracted from peripheral blood leukocytes with the Paxgene Blood RNA Kit (Qiagen) and reverse-transcribed (RT) into complementary DNA (cDNA) using the Superscript III Kit (Invitrogen). RT-PCR primers were in exon 33/34 (gaa cct ggc tta cca ggt ata) and in exon 42 (agg acc ttc tgg acc tgg tag). A PCR product bridging the deletion breakpoint was amplified with primers in intron 37 (aag cac cac ata ttc aag ttt c) and intron 41 (aac ttg cat gtt aat tca gac c) from control and patient DNA. Direct sequencing of all PCR products was carried out as described (Nozu et al. 2014b).

RepeatMasker analyses were performed with the sensitive cross-match search engine (v. 4) available at http://www.repeatmasker.org/. Prediction of RNA-binding proteins that may contact consensus binding sites in the bipartite exon was carried out with RBPmap (Paz et al. 2014). RNA secondary structures were predicted with overlapping sequences encompassing the new exon using a free energy minimization algorithm implemented in RNAStructure (Mathews 2006). The intrinsic strength of cryptic splice sites activated by genomic deletion was estimated by computing their maximum entropy scores (Yeo et al. 2004) and was compared to the score means for authentic counterparts of mutation-induced aberrant splice sites (Vorechovsky 2006; Buratti et al. 2007a). Auxiliary splicing sequences across the deletion breakpoint were examined against hexamer lists that were previously derived from computational and/or experimental studies of splicing enhancers or silencers (Fairbrother et al. 2004; Wang et al. 2004; Goren et al. 2006; Smith et al. 2006; Ke et al. 2010). The probability of upaired (PU) values, which serve as a useful measure of single-strandedness and correlate with functional splicing motifs, were computed as described (Hiller et al. 2007) using the new exon and 100 base-pairs (bp) of flanking intronic sequences as an input.

Results

The proband was a 10-year old girl without a family history of kidney disease. She was identified by chance hematuria and proteinuria. Her kidney biopsy showed characteristic basket-weave changes of the glomerular basement membrane (Kashtan 1999) visualized by electron microscopy (data not shown), leading to the diagnosis of a typical Alport syndrome. However, next-generation sequencing of *COL4A3*, *COL4A4* and *COL4A5* with a HaloPlex target system failed to detect any pathogenic variants. The MLPA carried out as the next diagnostic step identified a heterozygous *de novo* deletion encompassing *COL4A5* exons 38 through 41 (Figure 1A). Reverse transcriptase (RT)-PCR followed by direct sequencing of cDNA showed that exons 38 to 41 were replaced by a 72-bp insertion of a new exon, which was fully included in mature transcripts and introduced a stop codon in the mRNA (Figure 1B,C). Direct sequencing of genomic DNA revealed that the new exon was bipartite, originating from intron 37 (33 bp) and intron 41 (39 bp), with deletion breakpoints at c.3373+6282 and c.3791-2599 (Figure 1D,E). The exon was surrounded by canonical AG and GT dinucleotides that characterize the vast majority of human introns (Figure 1E).

RepeatMasker analysis of introns containing the deletion breakpoints revealed that the 3’ splice site and the 5’ part of the new exon were derived from an inactive antisense LINE-1 (L1) element (Figure 1F). This ancient L1ME copy also harbored the polypyrimidine tract and a high-score branch site of the bipartite exon (Figure 1E,F). The predicted branch point was located 20 bp upstream of the 3’ splice site, which was within the optimal distance previously estimated between 18 and 23 bp (Luukkonen et al. 1997; Chua et al. 2001). However, recognizable L1ME sequences did not extend into the 3’ portion of the intron 37-derived exon segment, nor were any repetitive elements detected in the intron 41-derived part of the new exon, including its 5’ splice site. Interestingly, a highly conserved AG dinucleotide at the 3’ splice site of the cryptic exon was absent in the L1ME consensus (Figure 1F). Sequence alignments of mammalian *COL4A5* genes showed that the AG dinucleotide was absent in rodent L1 orthologs, but was present in all primates, except for *Otolemur garnettii* (Figure 2). Assuming previously published estimates of the evolutionary age (Schmitz and Brosius, 2011), this indicated that mutations required for the bipartite exon activation in our patient took place in primitive primates roughly 85 million years ago. Finally, comparison of the exonized *COL4A5* L1 repeat with the exon-intron structure of ~150 existing human L1-derived exons (Sela et al. 2007) failed to uncover any exons with a splice site at the same L1 position, revealing an entirely new type of deletion-induced L1 exonization.

Accurate pre-mRNA splicing depends on proper local folding of nascent transcripts, which can facilitate or inhibit splice-site usage (Buratti et al. 2007b; Warf et al. 2010). Interestingly, secondary structure predictions of sequences surrounding the deletion breakpoint revealed reverse complementarity between the end of shortened intron 37 and the remaining part of intron 41, which might facilitate formation of a putative stem, bringing the silent splice sites into proximity (Figure S1). This structure could support cross-exon interactions that promote exon recognition, such as those involving serine/arginine-rich proteins or other RNA-binding factors predicted to bind motifs flanking the deletion breakpoint (Figure 1G). The deletion created a new UGCU motif, which contributes to optimal binding sites of at least four splicing factors (Figure 1G), including the well-characterized YGCY site of muscle blind-like proteins 1-3 (MBNLs) (Taliaferro et al. 2016 and references therein). The minimal binding site of MBNLs (underlined) was in a predicted single-stranded conformation (Figure 1G). The importance for secondary structure in this exonization event is supported also by the absence of any predicted splicing enhancer hexamers created by the deletion breakpoint (Table 1).

In contrast to auxiliary sequences, the intrinsic strength of both splice sites of the L1 exon was relatively high and was above the average for the 5’ splice site (Table 2). This splice site is flanked by a predicted stable helix, which may extend up to canonical base-pairing between the U1 small nuclear RNA and 5’ splice site at core intron positions -2 through +4 (Figure 3A). This interaction is probably stabilized by two pseudouridines at position +3 and +4, which may promote base-stacking (Davis 1995). As the adjacent position +5 is usually occupied by a conserved guanine and its point mutations are particularly vulnerable to cryptic 5’ splice site activation (Buratti et al. 2007a), selection of a weaker, L1-derived 3’ splice site may have been driven by the stronger 5’ splice site. Formation of the stable stem was supported by a very low probability of unpaired (PU) values computed across this region (Figure 3B).

Discussion

To our knowledge, this case represents the first *de novo* activation of a bipartite exon in genetic disease. Such 'dual-intron' exonization has not been reported even for fusion transcripts that often arise in cancer cells as a result of genomic rearrangements (Professor N. C. Cross, personal communication). Importantly, without analyzing RNA products, the *COL4A5* deletion alone would be expected not to alter the reading frame (as in Ensembl transcript *COL4A5-001*), omitting only three helix repeats from the entire collagen chain. In-frame deletions have been associated with less severe ultrastructural kidney damage and/or XLAS phenotypes (Mazzucco et al. 1998; Nozu et al. 2014b), although dominant negative effects of the mutated allele cannot be excluded. Together, these results highlight the importance of characterizing aberrant transcripts for accurate prognosis of this patient and hereditary disorders in general, with potentially important implications for their management. Assuming that the nonsense transcript identified this case (Figure 1) is indeed a cause of the typical female XLAS, repression of the new L1 exon by mono- or bipartite splice-switching oligonucleotides should increase the fraction of in-frame transcripts. This approach could ameliorate the phenotype in a manner similar to antisense-induced exon skipping in muscular dystrophy in the future (Aartsma-Rus 2010).

Symptomatic intragenic deletions often involve transposons (Guo et al. 2013) (Figure 1F). L1s are the most abundant autonomous retrotransposons, with >0.5 million copies in the human genome (Lander et al. 2001). The majority of L1s are inactive, with only a hundred of full-length copies capable of retrotransposition per genome (Brouha et al. 2003). *COL4A5* contains a single, potentially 'hot' L1 in intron 1 (Mir et al. 2015), but the exonized L1ME copy has been inactive for a very long time as it is interrupted by younger elements, including a DNA transposon Tigger3 (Figure 1F). The age of the primate-specific Tigger3 family was estimated between 54 and 67 million years (Pace et al. 2007).

As exemplified by the exonized *COL4A5* L1 (Figure 1F), intronic L1 insertions are biased towards the antisense orientation relative to mRNAs, both in humans and mouse (Sela et al. 2007), possibly as a result of selective pressure acting to prevent interference between the L1 and the host gene transcription. However, LINE repeats do not exhibit preferential exonization orientation (Sela et al. 2007). Also, exonized sequences derived from L1 elements usually comprise the whole exon rather than only 3’ or 5’ splice site (Sela et al. 2007), suggesting that they can be recognized by the spliceosome without assistance from flanking unique sequences. By contrast, the L1 exonization in *COL4A5* was limited to the 5’ end of the repeat, with recognizable L1 sequences occupying only the 5’ part of the fusion exon (Figure 1). Nevertheless, the overall exonization potential of intronic L1 repeats in evolution appears to be similar to other transposed elements (0.07%), but about 3x less than for *Alu*s (0.2%) (Sela et al. 2007), highlighting the unique character of the XLAS mutation.

RNA structure is an important modifier of exon selection, activating or inhibiting splice-site usage and also binding affinities of numerous proteins in the spliceosome (Warf et al. 2010). Recent work suggested that local RNA structure limits rather than promotes binding of MBNL1 (Taliaferro et al. 2016), one of the candidate proteins that may bind new sequence motifs created by the deletion and (de-)stabilize the stem-loop (Figure 1G). The predicted helix at the 5’ splice site could promote this site in the new genomic context (Figure 3A). The helix is capped by an atypical UAUA tetraloop that was previously found in an RNase P ribozyme (Harris et al. 2001). Important precedents for this scenario include the 5’ splice sites of *SMN2* exon 7 and *MAPT* exon 10, which are regulated by the stability adjacent stem-loops (Donahue et al. 2006; Singh et al. 2007). The *MAPT* exon 10 hairpin interacts with the DDX5 helicase (Kar et al. 2011). Finally, we cannot exclude that inclusion of the bipartite L1 exon in the *COL4A5* mRNA is immune to the deletion-mediated disruption in the transcription elongation rate (Han et al. 2004), which could affect RNA polymerase II processivity and kinetics of the spliceosome assembly (Luco et al. 2011 and references therein). Identification of key RNA-RNA and RNA-protein interactions that promoted the birth of the cryptic L1 exon will require experimental testing of these hypotheses in the future.

In conclusion, our case report demonstrates the astonishing versatility of intragenic deletions and transposed elements in shaping the new exon-intron structure, expanding the repertoire of currently known L1-mediated morbidities (Narita et al. 1993; Chen et al. 2005; Vorechovsky 2010). Our XLAS case also highlights the perilous inadequacy of predicting phenotypic severity of Mendelian disorders from DNA changes alone. As with exonized mammalian-wide interspersed repeats (Kralovicova et al. 2015), future systematic analyses of L1-derived 3’ splice sites should help characterize RNA interactions that facilitate their recognition by the spliceosome. Finally, our results suggest that the fraction of disease-causing intragenic deletions that affect RNA processing could be much larger than anticipated and that such cases may provide valuable exon selection models for future studies.

Ethical compliance

All procedures were reviewed and approved by the Institutional Review Board of Kobe University School of Medicine. Informed consent was obtained from proband's parents.

Conflict of interest: The authors have nothing to disclose.

Supporting information includes Figure S1 in a single pdf file:

Figure S1 Complementarity of the 5' and 3' parts of the bipartite exon

Acknowledgement

This study was supported by a grant from the Ministry of Health, Labour, and Welfare (Japan) for Research on Rare Intractable Diseases in the Kidney and Urinary Tract (H24-nanchitou (nan)-ippan-041 to Kazumoto Iijima) in the "Research on Measures for Intractable Diseases" Project, and a Grant-in-Aid for Scientific Research (KAKENHI) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (Subject ID: 25893131 to KN and 26293203 to KI). JK was supported by a Bloodwise grant (12060) to IV.

References

Aartsma-Rus A. 2010. Antisense-mediated modulation of splicing: therapeutic implications for Duchenne muscular dystrophy. RNA Biol. 7:453-461.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV and others. 2003. Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA 100:5280-5285.

Buratti E, Chivers MC, Hwang G, Vorechovsky I. 2011. DBASS3 and DBASS5: databases of aberrant 3' and 5' splice sites in human disease genes. Nucleic Acids Res. 39:D86-91.

Buratti E, Chivers MC, Kralovicova J, Romano M, Baralle M, Krainer AR and others. 2007a. Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic Acids Res. 35:4250-4263.

Buratti E, Dhir A, Lewandowska MA, Baralle FE. 2007b. RNA structure is a key regulatory element in pathological *ATM* and *CFTR* pseudoexon inclusion events. Nucleic Acids Res. 35:4369-4383.

Busslinger M, Moschonas N, Flavell RA. 1981. Beta+ thalassemia: aberrant splicing results from a single point mutation in an intron. Cell 27:289-298.

Chen JM, Stenson PD, Cooper DN, Ferec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum. Genet. 117:411-427.

Chua K, Reed R. 2001. An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. Mol. Cell. Biol. 21:1509-1514.

Corvelo A, Hallegger M, Smith CW, Eyras E. 2010. Genome-wide association between branch point properties and alternative splicing. PLoS Comput Biol 6:e1001016.

Davis DR. 1995. Stabilization of RNA stacking by pseudouridine. Nucleic Acids Res. 23:5020-5026.

Donahue CP, Muratore C, Wu JY, Kosik KS, Wolfe MS. 2006. Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. J. Biol. Chem. 281:23302-23306.

Fairbrother WG, Chasin LA. 2000. Human genomic sequences that inhibit splicing. Mol. Cell. Biol. 20:6816-6825.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA and others. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons.

Nucleic Acids Res. 32:W187-190.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I and others. 2006. Comparative analysis identifies exonic splicing regulatory sequences: The complex definition of enhancers and silencers. Mol. Cell 22:769-781.

Guo W, Zheng B, Cai Z, Xu L, Guo D, Cao L and others. 2013. The polymorphic *AluYb8* insertion in the *MUTYH* gene is associated with reduced type 1 protein expression and reduced mitochondrial DNA content. PLoS One 8:e70718.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429:268-274.

Harris JK, Haas ES, Williams D, Frank DN, Brown JW. 2001. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. RNA 7:220-232.

Hiller M, Zhang Z, Backofen R, Stamm S. 2007. Pre-mRNA secondary structures influence exon recognition. PLoS Genet. 3:e204.

Kar A, Fushimi K, Zhou X, Ray P, Shi C, Chen X and others. 2011. RNA Helicase p68 (DDX5) Regulates tau Exon 10 Splicing by Modulating a Stem-Loop Structure at the 5' Splice Site. Mol. Cell. Biol. 31:1812-1821.

Kashtan CE. 1999. Alport syndrome. An inherited disorder of renal, ocular, and cochlear basement membranes. Medicine (Baltimore). 78:338-360.

Ke S, Chasin LA. 2010. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. Genome Biol. 11:R84.

King K, Flinter FA, Nihalani V, Green PM. 2002. Unusual deep intronic mutations in the *COL4A5* gene cause X linked Alport syndrome. Hum. Genet. 111:548-554.

Kralovicova J, Patel A, Searle M, Vorechovsky I. 2015. The role of short RNA loops in recognition of a single-hairpin exon derived from a mammalian-wide interspersed repeat. RNA Biol. 12:54-69.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J and others. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. 2011. Epigenetics in alternative pre-mRNA splicing. Cell 144:16-26.

Luukkonen BG, Seraphin B. 1997. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. EMBO J. 16:779-792.

Mathews DH. 2006. RNA secondary structure analysis using RNAstructure. Curr. Protoc.

Bioinformatics 12:12.16.

Mazzucco G, Barsotti P, Muda AO, Fortunato M, Mihatsch M, Torri-Tarelli L and others. 1998. Ultrastructural and immunohistochemical findings in Alport's syndrome: a study of 108 patients from 97 Italian families with particular emphasis on *COL4A5* gene mutation correlations. J. Am. Soc. Nephrol. 9:1023-1031.

Mir AA, Philippe C, Cristofari G. 2015. euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res. 43:D43-47.

Narita N, Nishio H, Kitoh Y, Ishikawa Y, Ishikawa Y, Minami R and others. 1993. Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. J. Clin. Invest. 91:1862-1867.

Nozu K, Iijima K, Ohtsuka Y, Fu XJ, Kaito H, Nakanishi K and others. 2014a. Alport syndrome caused by a *COL4A5* deletion and exonization of an adjacent *AluY*. Mol. Genet. Genomic Med. 2:451-453.

Nozu K, Vorechovsky I, Kaito H, Fu XJ, Nakanishi K, Hashimura Y and others. 2014b. X-linked Alport syndrome caused by splicing mutations in *COL4A5*. Clin. J. Am. Soc. Nephrol. 9:1958-1964.

Oka M, Nozu K, Kaito H, Fu XJ, Nakanishi K, Hashimura Y and others. 2014. Natural history of genetically proven autosomal recessive Alport syndrome. Pediatr. Nephrol. 29:1535-1544.

Pace JK, 2nd, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res. 17:422-432.

Pagani F, Buratti E, Stuani C, Bendix R, Dork T, Baralle FE. 2002. A new type of mutation causes a splicing defect in *ATM*. Nat. Genet. 30:426-429.

Paz I, Kosti I, Ares M, Jr., Cline M, Mandel-Gutfreund Y. 2014. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 42:W361-367.

Schmitz J, Brosius J. 2011. Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie 93:1928-1934.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, 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 Biol. 8:R127.

Singh NN, Singh RN, Androphy EJ. 2007. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res. 35:371-389.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum. Mol. Genet. 15:2490-2508.

Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS and others. 2016. RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol. Cell 64:294-306.

Vorechovsky I. 2006. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucleic Acids Res. 34:4630-4641.

Vorechovsky I. 2010. Transposable elements in disease-associated cryptic exons. Hum. Genet. 127:135-154.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. Cell 119:831-845.

Warf MB, Berglund JA. 2010. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. 35:169-178.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. 11:377-394.

Tables

Table 1 Auxiliary splicing motifs created by the intronic fusion in *COL4A5*

New hexamer	Assignment (Ke et al. 2010)	PESE (Ke et al. 2010)	PESS (Ke et al. 2010)	RESCUE-ESE (Fairbrother et al. 2004)	ESS (Wang et al. 2004)	ESR (Goren et al. 2006)	ESE (Smith et al. 2006)
CAGTGC	Silencer	-	-	-	-	-	-
AGTGCT	Silencer	-	-	-	-	+	-
GTGCTG	Neutral	-	-	-	-	-	-
TGCTGT	Neutral	+	-	-	-	-	-
GCTGTT	Neutral	+	-	-	-	+	-

Legend: The underlined portions of hexamer motifs are derived from intron 37, the remaining part from intron 41. Hexamers found in the motif list referenced at the top are denoted by a plus sign. Abbreviations for the auxiliary splicing motifs (columns 3-8) are explained in cited references.

Table 2 The intrinsic strength of splice sites of the L1 exon in *COL4A5*

	Splice site	Maximum entropy score
COL4A5 L1 exon	3’	6.1
	5’	9.0
Authentic splice sites	3’	7.9
	5’	7.6

Figure legends

Figure 1 Genomic deletion activating a bipartite *COL4A5* exon in Alport syndrome

A, Identification of a heterozygous deletion by multiplex ligation-dependent probe amplification (MLPA) analysis of *COL4A5* exons. Exons with signal intensities at ~0.5 are denoted by a horizontal bar. Y-axis, normalized MLPA values. Both parents showed a normal MLPA pattern and no evidence for mosaicism (data not shown). **B,** RT-PCR of control (C) and patient (P) total RNA samples. M, DNA size marker (bp). PCR products are shown schematically to the right. The new boundary is denoted by a vertical red line, the new exon is in blue and canonical exons are numbered. Amplification primers (arrows) were located in exon 33/34 and exon 42. **C,** Sequence chromatogram of the aberrant cDNA product revealing a cryptic exon (blue bar). The L1 homology region (boxed) extends into the 5' portion of the bipartite exon; stop codon is underlined. **D,** PCR product amplified across the deletion breakpoint from control (C) and patient (P) DNA using primers in the indicated introns. **E,** Sequence chromatogram of the corresponding fragment. For legend, see panel C. The polypyrimidine tract (PPT, grey bar) and the branch point adenine (BP, in yellow) of the new exon were predicted by a support vector machine (SVM) algorithm, with a SVM score of 0.81 (Corvelo et al. 2010). Conserved dinucleotides at new splice sites are in red boxes. **F,** Summary of transposed elements across the centromeric deletion breakpoint (*upper panel*) and the alignment of the L1ME in *COL4A5* intron 37 with a L1ME consensus (*lower panel*). Mutation creating the AG dinucleotide in the *COL4A5* L1ME is in red. **G,** Putative interactions between RNA-binding proteins and sequence motifs flanking the deletion breakpoint, as predicted by the RBPmap (Paz et al. 2014).

Figure 2 Sequence alignment of mammalian *COL4A5* orthologs across the L1 exon-derived 3' splice site

Conserved AG dinucleotide is in red, predicted branch point adenine in yellow. Alignment was created with full genomic reference sequences using Clustal Omega (v. 1.2.2). -, deletion, ., not determined.

Figure 3 Predicted stem-loop flanking the 5' splice site of the bipartite exon

A, Base-pairing interactions between the U1 small nuclear RNA and the 5' splice site. L1 exon (highlighted in purple) is in upper case, intron 41 sequences are in lower case, Ψ, pseudouridine. **B,** PU values across this region. The predicted stem-loop is denoted by a horizontal black bar and unpaired nucleotides of the stem-loop by closed circles.

Fig. 1

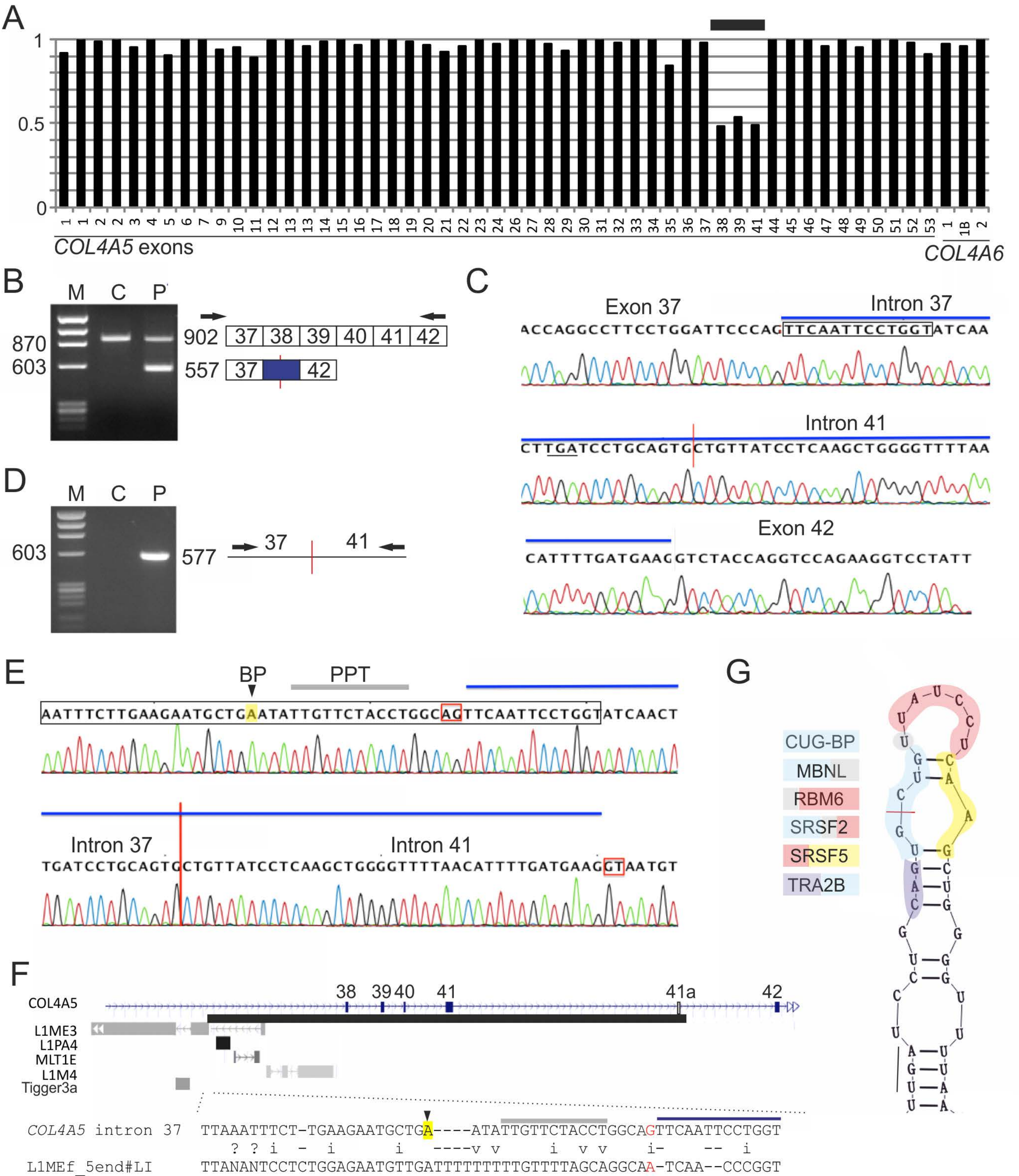
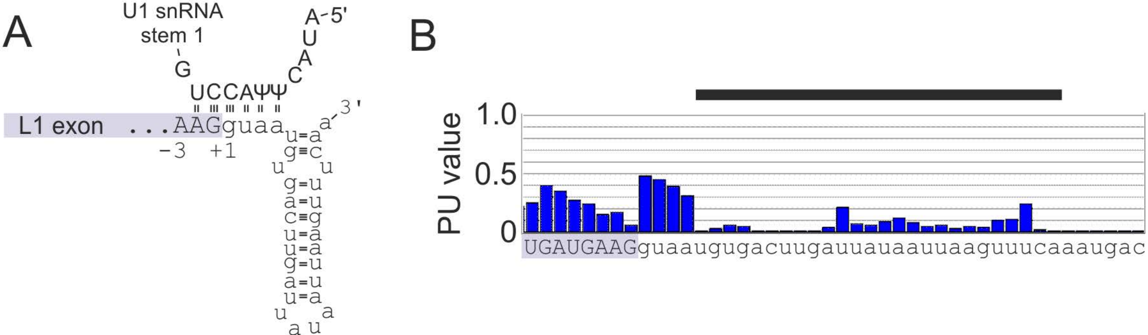

A
1
0.5
0
1 1 2 2 3 4 5 6 7 9 10 11 12 13 13 14 15 16 17 18 19 20 21 22 23 24 26 27 28 29 30 31 32 33 34 35 36 37 38 39 41 44 45 46 47 48 49 50 51 52 53
COL4A5 exons
1 1B 2
COL4A6
B
M C P
870
603
902 37 38 39 40 41 42
557 37 42
C
Exon 37
Intron 37
ACCAGGCCTTCCTGGATTCCCAGTTCAATTCCTGGTATCAA
Intron 41
CTTGATCCTGCAGTGCTGTTATCCTCAAGCTGGGGTTTTAA
Exon 42
CATTTTGATGAAGGTCTACCAGGTCCAGAAGGTCCTATT
D
M C P
603
577
37
41
E
BP
PPT
AATTTCTTGAAGAATGCTGAATATTGTTCTACCTGGCAGTTCAATTCCTGGTATCAACT
Intron 37
Intron 41
TGATCCTGCAGTGCTGTTATCCTCAAGCTGGGGTTTTAACATTTTGATGAAGGTAATGT
F
COL4A5
38
39
40
41
41a
42
L1ME3
L1PA4
MLT1E
L1M4
Tigger3a
COL4A5 intron 37 TTAAATTTCT-TGAAGAATGCTGA----ATATTGTTCTACCTGGCAGTTCAATTCCTGGT
? ? i - i i ----v v i v v i- -- i
L1MEf_5end#LI TTANANTCCTCTGGAGAATGTTGATTTTTTTTTGTTTTAGCAGGCAA-TCAA--CCCGGT
G
CUG-BP
MBNL
RBM6
SRSF2
SRSF5
TRA2B
U A U C C U U G U C G U G A C G U C U U A G U U A C A G C U G G U G G U U U A A

Fig. 2

```
Homo sapiens             CTATTAAATTTCT-TG--AAGAATGCTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Pan troglodytes          CTATTAAATTTCT-TG--AAGAATGCTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Gorilla gorilla          CTATTAAATTTCT-TG--AAGAATGCTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Pongo abelii             CTATTAAATTTCT-TG--AAGAATACTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Nomascus leucogenys      CTATTAAATTTCT-TG--AAGAATGCTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Chlorocebus sabaeus      CTATTAAATTTCT-TG--AAGAATACTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Macaca mulatta           CTATTAAATTTCT-TG--AAGAATACTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Papio anubis             CTATTAAATTTCT-TG--AAGAATACTGAATATTGTTCTACCTGGCAGTTCAATTCCTG
Callithrix jacchus       CTATTAAATTTCTTTG--AATAATATTGAATATTGTTCTAGCTGGCAGTTCAATTCCTG
Tarsius syrichta         ......................GTGCCAAATATTGTTTCAGCAGGCAGTTCAATTCAGA
Microcebus murinus       CTATTAAATTTCT-TG--AAGAGTGCTGAATATTGTTCTA---GGCAGTTCAATTCCTG
Otolemur garnettii       CTATTAAATTTCT-TG--AAGAGTGCTGAATATTATTCTAGTAGGCCATAAATCACCTG
Mus musculus             CTGTCAAATTTCT-TTATAAAAACGTTGACTA---TTCTATCAGGTCATCCATTTTTTG
Rattus norvegicus        CTCTTAAATTTCT-TTGTAAAAATGTTGACTA---TTATATCAAGTCATTCATTTCCTG
```

Fig. 3


A
U1 snRNA
stem 1
G
UCCAΨΨ
C
A
U
A-5'
3'
L1 exon ...AAGguaa
a
−3
+1
u=a
g≡c
u u
g=u
a=u
c≡g
u=a
u=a
g=u
a=u
u=a
u a
a u
B
PU value
1.0
0.5
0
UGAUGAAGguaaugugacuugauuauaauuaaguuucaaaugac